Statistical analysis of longitudinal randomized clinical trials with missing data: a comparison of approaches

Royes Joseph

A thesis submitted for the degree of Doctor of Philosophy

March 2015

Research Institute for Primary Care and Health Sciences

Keele University

Keele University

Declaration Part 1

Declaration

SUBMISSION OF THESIS FOR A RESEARCH DEGREE

Degree for which thesis being submitted: PhD

Title of thesis: Statistical analysis of longitudinal randomized clinical trials with missing data: a comparison of approaches

This thesis contains confidential information and is subject to the protocol set down for the submission and examination of such a thesis. No

Date of submission: 05 November 2014 Original registration date: 01 October 2011

Name of candidate: Royes Joseph
Research Institute: Primary Care and Health Sciences
Name of Lead Supervisor: Dr Martyn Lewis

I certify that:

(a) The thesis being submitted for examination is my own account of my own research
(b) My research has been conducted ethically. Where relevant a letter from the approving body confirming that ethical approval has been given has been bound in the thesis as an Annex
(c) The data and results presented are the genuine data and results actually obtained by me during the conduct of the research
(d) Where I have drawn on the work, ideas and results of others this has been appropriately acknowledged in the thesis
(e) Where any collaboration has taken place with one or more other researchers, I have included within an 'Acknowledgments' section in the thesis a clear statement of their contributions, in line with the relevant statement in the Code of Practice (see Note overleaf).
(f) The greater portion of the work described in the thesis has been undertaken subsequent to my registration for the higher degree for which I am submitting for examination
(g) Where part of the work described in the thesis has previously been incorporated in another thesis submitted by me for a higher degree (if any), this has been identified and acknowledged in the thesis
(h) The thesis submitted is within the required word limit as specified in the Regulations

Total words in submitted thesis (including text and footnotes, but excluding references and appendices): 64624

Signature of candidate ……Date: 23/02/2015

Abstract

Objectives

Missing data represent a source of bias in randomized clinical trials (RCTs). This thesis focuses on pragmatic RCTs with missing continuous outcome data and evaluates the use and appropriateness of current methods of analysis.

Methods

This thesis consists of three parts. First, a systematic review examined practices relating to missing data in published RCTs. Second, a simulation study compared the performance of various methods for handling missing data in a number of plausible trial scenarios. Finally, an empirical evaluation of two pragmatic RCTs investigated the use of a reminder process to inform whether missingness is likely to be non-ignorable.

Results

The majority of 91 trials in the systematic review adopted a form of single imputation, such as last observation carried forward (LOCF) for dealing with missing data. Mixed-effects model for repeated measures (MMRM) and/or multiple imputation (MI) were limited to eight trials. Sensitivity analyses were infrequently and inappropriately used, and insufficiently reported.

In the simulation study, LOCF yielded biased estimates of treatment effect in most scenarios, irrespective of missing data mechanisms. All methods, except LOCF, yielded unbiased estimates for scenarios of equal dropout rate and same direction of dropout in both treatment groups. MMRM and MI were more robust to bias than complete-case and LOCF-based analyses.

In the empirical study, the evaluation using reminder responses indicated the possibility of biased MMRM estimation in one trial and unbiased MMRM estimation in the other.

Conclusion

CCA and LOCF-based analysis should be disregarded in favour of methods such as MMRM and MI-based analysis. The proposed reminder approach can be used to assess the robustness of the missing at random (MAR) assumption by checking expected consistency in MAR-based estimates. If the results deviate, then analyses incorporating a range of plausible missing not at random assumptions are advisable, at least as sensitivity tests for the evaluation of treatment effect.

Table of contents

Declaration ..... **i**

Abstract ..... **ii**

Table of contents ..... **iv**

List of tables ..... **xi**

List of figures ..... **xiii**

List of abbreviations ..... **xvi**

List of publications ..... **xviii**

Acknowledgements ..... **xix**

Chapter 1: Introduction **1**

1.1 The present study: research issues 1

1.2 Outline of thesis 3

Chapter 2: Background **6**

2.1 Introduction 6

2.2 The problem of missing data in clinical trials 6

2.3 Missing data: theoretical framework 11

2.3.1 Missing data patterns 12

2.3.2 Missing data mechanism 13

2.3.3 Identification of the missing data mechanism 15

2.4 Methods for handling incomplete continuous data 15

2.4.1 Description of candidate approaches to analysing longitudinal RCT data with missing values 16

2.4.2 Choice of approach to handling missing data: existing literature on comparisons of the approaches 27

2.4.3 Sample size calculation in anticipation of dropouts 34

2.5 Guidance on prevention and handling of missing data in RCTs 35

2.6 Rationale of the thesis 36

2.7 Conclusion 38

Chapter 3: Systematic review 40

3.1 Introduction 40

3.2 Background 40

3.2.1 RCTs in MSCs 44

3.2.2 ITT analysis and missing outcome data in trials in MSCs 45

3.2.3 Handling of missing outcome data in trials in MSCs 46

3.3 Objectives 46

3.4 Methods 47

3.4.1 Selection of studies 47

3.4.2 Data extraction and management 50

3.5 Results 55

3.5.1 Characteristics of included trials 55

3.5.2 Dropouts 57

3.5.3 Analysis strategy and loss to analysis 60

3.5.4 Handling of dropouts: imputation strategy 63

3.5.5 Sensitivity analysis and cautionary notes on missing data 65

3.6 Discussion 67

3.6.1 Overall summary 67
3.6.2 Quality of reporting 68
3.6.3 Importance of collecting data on all randomized subjects 69
3.6.4 Power calculation in anticipation of dropouts 70
3.6.5 Dropout rate 71
3.6.6 Analysis strategy 74
3.6.7 Baseline comparison 76
3.6.8 Handling missing data 77
3.6.9 Sensitivity analysis 81
3.7 Limitations and generalizability 82
3.8 Conclusion 82
Chapter 4: Simulation study: an overview of design 84
4.1 Introduction 84
4.2 Background 85
4.3 Simulation procedure 86
4.3.1 Step 1: Generating complete datasets 86
4.3.2 Step 2: Generating missing data 93
4.3.3 Step 3: Imputation and analysis methods 97
4.4 Measures of performance 99
4.4.1 Bias 99
4.4.2 Overall accuracy of the estimate 100
4.4.3 Coverage of confidence interval 100

4.4.4 Average width of confidence interval 101
4.4.5 Statistical power 101
4.5 Summary of simulation scenarios 102
4.6 Discussion and conclusion....104
Chapter 5: Simulation study - findings 1....109
5.1 Introduction....109
5.2 Bias and precision 112
5.2.1 Bias and RMSE under MCAR....112
5.2.2 Bias and RMSE under MAR dependent on baseline value (MAR-B)....116
5.2.3 Bias and RMSE under MAR dependent on last observed value (MAR-L) 119
5.2.4 Bias and RMSE under MNAR....123
5.3 Confidence interval coverage and width....127
5.3.1 CI coverage and width under MCAR....128
5.3.2 CI coverage and width under MAR-B....131
5.3.3 CI coverage and width under MAR-L....134
5.3.4 CI coverage and width under MNAR....138
5.4 Statistical power to detect the true difference....141
5.4.1 Statistical power under MCAR....142
5.4.2 Statistical power under MAR-B....144
5.4.3 Statistical power under MAR-L....146
5.4.4 Statistical power under MNAR....148
5.5 Summary of findings....150

Chapter 6: Simulation study: findings 2.....152

6.1 Introduction.....152

6.2 Effect of trajectory pattern and size of treatment effect on inferences from the missing data handling approaches.....153

6.3 Comparison of two strategies for handling baseline data in an MMRM analysis... 159

6.3.1 Effects on overall accuracy.....160

6.3.2 Effects on the coverage of 95% CI 160

6.3.3 Effects on the observed power 163

6.4 Effect of sample size on power under different missing data mechanisms: a comparison of missing data handling approaches.....165

6.4.1 When the desired power was 90% in the absence of missing data.....166

6.4.2 When the desired power was 80% in the absence of missing data.....171

6.5 Summary of findings.....174

6.6 Overall summary of findings from simulation studies.....176

Chapter 7: An empirical evaluation of the impact of missing data on treatment effect: analysis of TATE and STarT Back trials.....178

7.1 Introduction.....178

7.2 Background 178

7.3 Reminder responses as proxies of non-responses 179

7.4 Methods.....181

7.5 The TATE trial.....184

7.5.1 Descriptive analysis of missing data 186

7.5.2 Analysis of the incomplete TATE trial – estimation of treatment effect at month 12 193

7.5.3 Summary and interpretation of findings 198

7.6 The STarT Back trial 201

7.6.1 Descriptive analysis of missing data 202

7.6.2 Analysis of STarT Back trial data – estimation of the treatment effect at month 12 209

7.6.3 Summary and interpretation of findings 213

7.7 Discussion 214

7.8 Conclusion 219

Chapter 8: Summary, discussion and conclusions 221

8.1 Introduction 221

8.2 Summary of findings 222

8.2.1 Summary of systematic review 222

8.2.2 Summary of simulation study 224

8.2.3 Summary of empirical evaluation 228

8.3 Discussion of the findings 230

8.3.1 The performance of incomplete data analysis methods for the estimation of treatment effect in RCTs 230

8.3.2 Choice between MMRM and MI-based analyses in an RCT 237

8.3.3 Strategy for handling baseline values with MMRM analysis 241

8.3.4 The benefits of sample size inflation to the effect of attrition on statistical power 242

8.3.5 Reminder data to investigate the appropriateness of MAR-based analyses. 246

8.4 Limitations and generalizability 248

8.5 Implications for practice 250

8.6 Future work 253

8.7 Conclusion 254

References 256

Appendices 272

List of tables

Table 2.1: Percent efficiency of MI estimation 22

Table 3.1: Classification of analysis strategy used in trial reports 53

Table 3.2: Description of the selected trials (n=91) 56

Table 3.3: Size of the trial 57

Table 3.4: Analysis strategy followed in the primary analysis. Data are counts (%) 60

Table 3.5: Description on analysis strategy provided in the 28 trial reports with classification 'partial ITT' 62

Table 3.6: Methods used to handle late dropouts, who had completed at least one follow-up assessment. Data are counts (%) 64

Table 3.7: Description of trials that performed a sensitivity analysis for missing data 66

Table 3.8: Recommendations 3–5 of the NAS report on missing data 70

Table 4.1: Correlation and SD matrices for simulation scenarios 91

Table 4.2: Calculated sample size under various covariance patterns 92

Table 4.3: Sample size used for study 3 – effect of sample size 92

Table 4.4: Planned cumulative dropout rate (%) 94

Table 4.5: Simulation scenarios under study 1 102

Table 4.6: Simulation scenarios under study 2 103

Table 4.7: Simulation scenarios under study 3 103

Table 4.8: Simulation scenarios under study 4 104

Table 6.1: The observed power with inflated sample size – desired power was 90% 169

Table 6.2: The observed power with inflated sample size – desired power was 80% 172

Table 7.1: Responders' status at follow-up assessments 186

Table 7.2: The observed pairwise correlation between variables in the actual dataset...192

Table 7.3: TATE - ANCOVA results at 12 months follow-up before and after LOCF imputation of missing values.....194

Table 7.4: TATE - MMRM results.....195

Table 7.5: TATE - ANCOVA results after MI imputation of missing values.....196

Table 7.6: Results from the modified dataset.....198

Table 7.7: The observed pairwise correlation between variables 208

Table 7.8: STarT Back - ANCOVA results before and after LOCF imputation of missing values.....209

Table 7.9: STarT Back - MMRM results.....210

Table 7.10: STarT Back - ANCOVA results after MI imputation of missing values 211

Table 7.11: Results from the modified dataset – RMDQ 212

Table 8.1: Summary of simulation results 226

Table 8.2: Statistical power for analysis methods (among the scenarios in which the methods yielded unbiased estimates of treatment effect).....245

List of figures

Figure 2.1: Missing data patterns 12

Figure 3.1: Identification of randomized trials from January 2010 to December 2011 50

Figure 3.2: The distribution of the 90 trials based on the percentage of dropouts 58

Figure 3.3: Dropout rate at primary endpoint by number of follow-ups 59

Figure 3.4: Dropout rate at the primary endpoint between arms (n=61) 59

Figure 3.5: Handling of missing data in trials with >10% late dropouts (n=40; detail was not reported in one trial) 65

Figure 3.6: Comparison of reviews in relation to percentage of trials with various levels of dropout rates 72

Figure 4.1: Schematic diagram to show the simulation procedures 87

Figure 4.2: Assumed means trajectories 89

Figure 5.1: Bias under MCAR 114

Figure 5.2: RMSE under MCAR 115

Figure 5.3: Bias under MAR-B 117

Figure 5.4: RMSE under MAR-B 118

Figure 5.5: Bias under MAR-L 121

Figure 5.6: RMSE under MAR-L 122

Figure 5.7: Bias under MNAR – in relation to (a) level of dropout rate (%) with a fixed strong correlation and moderate SD; (b) data variability and correlation between the repeated assessments with 30% dropout rate 125

Figure 5.8: RMSE under MNAR – in relation to (a) level of dropout rate (%) with a fixed strong correlation and moderate SD; (b) data variability and correlation between the repeated assessments with 30% dropout rate 126

Figure 5.9: CI coverage under MCAR 129

Figure 5.10: Average width of the 95% CI under MCAR 130

Figure 5.11: CI coverage under MAR-B 132

Figure 5.12: Average width of the 95% CI under MAR-B 133

Figure 5.13: CI coverage under MAR-L 135

Figure 5.14: Average width of the 95% CI under MAR-L 136

Figure 5.15: CI coverage under MNAR – in relation to (a) level of dropout rate (%) with a fixed strong correlation and moderate SD; (b) data variability and correlation between the repeated assessments with 30% dropout rate 139

Figure 5.16: Average width of CI under MNAR – in relation to (a) level of dropout rate (%) with a fixed strong correlation and moderate SD; (b) data variability and correlation between the repeated assessments with 30% dropout rate 140

Figure 5.17: Statistical power under MCAR 143

Figure 5.18: Statistical power under MAR-B 145

Figure 5.19: Statistical power under MAR-L 147

Figure 5.20: Statistical power under MNAR – in relation to (a) level of dropout rate (%) with a fixed strong correlation and moderate SD; (b) data variability and correlation between the repeated assessments with 30% dropout rate 149

Figure 6.1: Effect of trajectory pattern and mean difference between groups over time on bias in estimate of treatment effect 155

Figure 6.2: Effect of trajectory pattern and mean difference between groups over time on RMSE of estimate 156

Figure 6.3: Effect of trajectory pattern and mean difference between groups over time on coverage of 95% CI 157

Figure 6.4: Effect of trajectory pattern and mean difference between groups over time on width of 95% CI 158

Figure 6.5: Coverage of 95% CI under various scenarios (MMRM – Mixed model with baseline-as-covariate; cLDA – Mixed model with baseline-as-outcome) 162

Figure 6.6: Statistical power under various scenarios (MMRM – Mixed model with baseline-as-covariate; cLDA – Mixed model with baseline-as-outcome) 164

Figure 6.7: Statistical power under different sample sizes (10% dropouts).....167

Figure 6.8: Statistical power under different sample sizes (30% dropouts).....168

Figure 7.1: Response rate (%) over time on outcome measures – (a) pain intensity, (b) PRTEE total score, and (c) SF12 score. 187

Figure 7.2: Observed mean profiles according to intervention groups and time at which participants lost to follow-up.....189

Figure 7.3: Response rate (%) over time on outcome variables – (a) RMDQ, (b) back pain intensity, and (c) SF12 (PCS & MCS) 204

Figure 7.4: Observed mean profile according to intervention groups and the time at which dropped out 205

List of abbreviations

ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
AT	As-treated
BOCF	Baseline observation carried forward
CCA	Complete-case analysis
CI	Confidence interval
CONSORT	CONsolidated Standards of Reporting Trials
EMA	European Medicines Agency
FCS	Full conditional specification
FIML	Full-information maximum likelihood
FITT	Full intention-to-treat
GEE	Generalized estimating equations
ICH	International Conference on Harmonisation
ITT	Intention-to-treat
JM	Joint modelling
LME	Linear mixed-effects
LOCF	Last observation carried forward
MAR	Missing at random
MAR-B	MAR dependent on baseline
MAR-L	MAR dependent on last observed value
MCAR	Missing completely at random
MCMC	Markov chain Monte Carlo
MCS	Mental component score of SF12
MDC	Minimum data collection
MI	Multiple imputation
MICE	Multiple imputation by chained equation
ML	Maximum likelihood
MMRM	Mixed-effects model for repeated measures
MNAR	Missing not at random
MSC	Musculoskeletal condition
MSE	Mean squared error

NAS	National Academy of Science
NRC	National Research Council
OR	Odds ratio
PCM	Primary care management
PCS	Physical component score of SF12
PITT	Partial intention-to-treat
PP	Per-protocol
PRTEE	Patient-rated tennis elbow evaluation
RCT	Randomized clinical trial
REML	Restricted maximum likelihood
RMDQ	Roland Morris Disability Questionnaire
RMSE	Root-mean-square error
SD	Standard deviation
SF 12	Short-form 12
TENS	Transcutaneous electrical nerve stimulation
wGEE	Weighted GEE
WOCF	Worst observation carried forward

List of publications

i. Royes Joseph, Julius Sim, Reuben Ogollah, Martyn Lewis. A systematic review finds variable use of the intention-to-treat principle in musculoskeletal randomized controlled trials with missing data. *Journal of Clinical Epidemiology* 2014 (DOI: 10.1016/j.jclinepi.2014.09.002).

ii. Royes Joseph, Julius Sim, Reuben Ogollah, Martyn Lewis. Evaluation of bias and precision in methods of analysis for pragmatic trials with missing outcome data: a simulation study. *Trials* 2013, 14(Suppl 1):P110 (Conference proceedings).

Acknowledgements

I would like to thank my supervisors Dr Martyn Lewis, Professor Julius Sim, and Dr Reuben Ogollah for their tremendous support and constructive feedback throughout this study. I am grateful for their invaluable time, advice and encouragement throughout my PhD.

My profound gratitude goes to the Keele University and the Research Institute for Primary Care and Health Sciences for providing the necessary grants and enabling environment that facilitated the timely completion of this thesis. I would also like to thank the TATE and STarT Back trials team for permitting the use of their data.

I would like to thank my family and friends for their support and encouragement.

I am extremely thankful to my wife Smitha Royes and my little son Johan for their encouragement and for being so patient with me all through the programme and, above all, God almighty, the giver of life and grace.

Chapter 1: Introduction

1.1 The present study: research issues

Randomized clinical trials (RCTs) play a vital role in assessing the efficacy and effectiveness of new interventions compared to a standard or control intervention. An intention-to-treat (ITT) strategy – whereby an analysis should be performed by including all study participants in the groups to which they were randomized, regardless of any departures from the original assigned group – serves to preserve the benefits of randomization, which is intended to ensure that differences in outcome observed between treatment groups are solely the result of the treatments, and to reduce the risk of selection bias. A true ITT analysis requires baseline and outcome measurements on all randomized patients. In practice, no matter how well designed and implemented a study, missing data are almost inevitable – particularly in pragmatic trials. Different degrees of data incompleteness in these trials can occur as measurements may be available only at baseline or may be missed for one or several follow-up time-points. In general there are three potential problems that arise from missing data; loss of efficiency, complication in data handling and analysis, and bias due to differences between the observed and unobserved data. Despite extensive literature on methods of handling missing data, it appears that many RCTs continue to be based on inappropriate statistical methods when dealing with missing data (Hollis & Campbell, 1999; Wood et al., 2004; Baron et al., 2005; Gravel et al., 2007; Fielding et al., 2008).

This thesis focused on RCTs with missing continuous outcome data, which are prone to dropouts due to their longitudinal nature. A particular focus is on pragmatic trials of musculoskeletal disorders in primary care (though much of the theory and findings relate more generally to other RCTs). Principally, the work aims to align current recommendations to the methods of analysis being used in practice, and evaluate the

appropriateness of current methods of analysis in respect of the validity in estimation of the true between-group treatment effect conditional on missing data. In order to meet the overall aim, the following objectives were identified.

i. To provide a general overview of the statistical methods to deal with missing data

ii. To investigate any divergence among researchers on acceptance of these methods for analysing missing data

iii. To review current practices being used in the analysis of RCTs in the presence of missing data

iv. To evaluate the impact of various missing data handling methods for the analysis of continuous outcomes in longitudinal clinical trials under various plausible conditions in a comprehensive manner using simulation studies

v. To propose and investigate how reminder responses – data that are retrieved by sending reminders to the initial non-responders – can be utilized to infer the nature of missing data and help inform appropriate analysis of the RCT dataset in order to reduce potential bias in treatment effect estimation

vi. To collectively appraise the various findings and provide recommendations on how to deal with missing data in RCTs

1.2 Outline of thesis

Chapter 2: Background

Chapter two provides an initial background to the issues associated with missing data and a summary of the current missing data literature. In particular, the review focuses on key issues relating to missing continuous outcome data and methods of handling missing data currently in use in longitudinal clinical trials. The methods include listwise deletion, single imputation (e.g. last observation carried forward method [LOCF]), multiple imputation (MI), and a maximum likelihood based approach (e.g. mixed-effects model for repeated measures [MMRM]) that can use all available data without imputation. This chapter discusses the advantages and disadvantages of these missing data methods based on a review of previous simulation studies that compared these methods. The chapter also discusses the limitation of these simulation studies and rationalizes the requirement for further study.

Chapter 3: Systematic review

In this chapter, I present a systematic review that examines current practice relating to ITT analysis and methods to handle missing data in published trials. Specifically, the review has the following objectives:

- To describe the extent of adherence to random allocation;
- To describe the extent of reported dropout;
- To summarize the frequency in use of different analytical methods used to handle missing data;
- To assess the use of sensitivity analyses used to assess the robustness of the primary analysis results to various missing data assumptions.

In this study, the review focuses on RCTs reported in five leading medical journals that mainly focus on research in musculoskeletal conditions.

Chapters 4–6: Simulation study

Chapter four details the methodology of a simulation study that investigates the performance of various methods for handling missing data in a longitudinal clinical trial with continuous outcome data, across a number of different scenarios. The methods include listwise deletion, LOCF, MI, and MMRM. The simulation study has the following objectives:

- To assess the relative performance of the missing data methods with respect to bias and accuracy of the estimate of treatment effect under various scenarios;
- To assess the relative performance of the missing data methods with respect to the coverage of confidence interval of the estimate of treatment effect at the nominal alpha level of 0.05 under various scenarios;
- To assess the relative performance of the missing data methods with respect to conditional loss of statistical power to detect the true treatment effect under various scenarios (given nominal power of 90%);
- To assess whether an increment in sample size in proportion to an expected dropout rate helps to achieve the required statistical power when using these missing data methods;
- To assess whether including the baseline measure as part of the response vector in an MMRM model has an advantage over including it as a covariate.

Chapters five and six present the simulation study results.

Chapter 7: Re-analysis of real incomplete longitudinal RCT datasets

Chapter seven presents the re-analysis of two pragmatic clinical trials that included a reminder process for non-responders. Here, I propose an approach that utilizes the reminder process as a proxy for missingness to assess the impact of missing data on the estimation of treatment effect, and the likely missing data mechanism.

Chapter 8: Discussion and conclusion

Chapter eight concludes with a detailed discussion and interpretation around the findings from all the chapters. I finish by providing a summary of my recommendations on how to deal with missing data in RCTs, and some thoughts on further research in this area.

Chapter 2: Background

2.1 Introduction

This chapter focuses on key issues relating to missing continuous outcome data and methods of handling the missing data currently in use in longitudinal clinical trials. This chapter also discusses the advantages and disadvantages of these missing data methods based on a review of available simulation studies that compared these methods. The chapter further discusses the limitation of these simulation studies and then conclude by identifying the need for a further study to compare these missing data techniques. Before describing the different missing data techniques, the definition and underlying theory of missing data are also presented.

2.2 The problem of missing data in clinical trials

Randomized clinical trials (RCTs) play a vital role in assessing the efficacy and effectiveness of new interventions compared to a standard or control intervention. Randomization in a clinical trial is intended to generate comparable groups of patients in terms of known and, more importantly, unknown factors that could be associated with the outcome of interest at the onset of the trial. That is, the method ensures at least theoretically, that both observed and unobserved baseline differences between the interventions are attributable to chance. After accounting for chance variations, the remaining differences can be attributed reliably to the interventions so long as other sources of bias have been eliminated. To provide an unbiased comparison of estimates of treatment effects, randomization alone is not sufficient and it is also important to obtain outcome measurements on all randomized patients. Therefore, the principal advantage of randomization is threatened when some outcome measurements are missing. As trials with missing data may not retain the balance of randomization, the basis for statistical inference is lost (Wright & Sim, 2003; Lewis & Machin, 1993) and there is no longer a

statistical rationale to guarantee lack of bias for the estimation of the parameter and its associated confidence interval – even if the study is assumed to be free of other risks of bias, such as non-masked evaluation.

The intention-to-treat (ITT) principle is widely recommended as the primary design and analysis strategy for clinical trials (Frangakis & Rubin, 1999; Feinman, 2009); this is mandatory for any confirmatory trial (Committee for Proprietary Medical Products, 2001; Food and Drug Administration, 2008). An ITT analysis is a pragmatic approach that may help to avoid bias in estimation of treatment effect occasioned by any study protocol violation after randomization, such as dropout of subjects, which may affect the baseline equivalence established by randomization (Schwartz & Lellouch, 2009). An ITT analysis corresponds to analysing groups exactly as randomized. Strictly, an ITT analysis should include all randomized subjects, regardless of their adherence with the eligibility criteria, the treatment they actually received, and subsequent withdrawal or loss to follow-up from treatment or deviation from the study protocol (Fisher et al., 1990). Accordingly, an ITT analysis includes all randomized subjects according to randomized treatment assignment. It ignores protocol deviations, non-compliance, withdrawal and anything that happens after randomization (Heritier et al., 2003; Kruse et al., 2002). An ITT analysis is generally explained as "*once randomized, always analysed*" (Wertz, 1993). Hollis and Campbell (1999) point out two purposes of an ITT approach: firstly, it maintains treatment groups that are similar apart from random variation, and secondly, it allows for non-compliance and deviations from policy by investigators. Thus, an ITT analysis reflects the practical clinical scenario. Therefore, an ITT analysis is most suitable for pragmatic trials, which measure the effectiveness of treatments in everyday practice (Hollis & Campbell, 1999).

It has been reported that investigators often refer to ITT to describe the analysis of all available subjects as randomized without considering the issue of missing data (Gravel et al., 2007). It is pointed out in the CONSORT (CONsolidated Standards Of Reporting

Trials) statement that a strict ITT analysis is often hard to achieve for two main reasons: (i) missing outcomes for some participants and (ii) non-adherence to the treatment protocol (Moher et al., 2010). Therefore, compliance with the ITT principle would necessitate complete follow-up of all randomized subjects for study outcomes and retention of randomized allocation grouping regardless of deviation from treatment protocol. Exclusion of participants, possibly in a non-random or informative way, raises great concerns about the validity of the study. Many reviews of RCTs concede that the ITT approach is often inadequately described and applied (Schulz et al., 1996; Hollis & Campbell, 1999; Kruse et al., 2002; Gravel et al., 2007) – the deviation from the true ITT was mostly linked in these reviews with missing data (Chapter 3). The International Conference on Harmonisation (ICH) guideline (Food and Drug Administration, 1997) states: "*no analysis should be considered complete unless the potential biases arising from these specific exclusions, or any others, are addressed*."

RCTs are generally longitudinal in nature – such that the outcome of interest is measured at more than one occasion – with a common schedule of measurements for all participants, but with a small number of measurement occasions. In a review of trial reports published between July and December 2001 in four major general medicine journals (*BMJ*, *JAMA*, *Lancet* and *New England Journal of Medicine*), among the 71 trial reports that were examined, 37 (52%) of them were with multiple follow-up assessments (Wood et al., 2004). Even though outcome data are observed in a longitudinal fashion, the primary focus of RCTs is often a specific time of measurement, usually the last – called the primary endpoint. The aim of these trials is usually limited to comparing the effect of two or more treatments at this specific time-point – i.e., estimating treatment effect at the primary endpoint (Verbeke & Molenberghs, 2005). Importantly, many musculoskeletal conditions (MSCs) necessitate long-term trials because they are chronic conditions, and this consequently results in a high number of patients being discontinued from the trial

prior to the primary endpoint (Kim, 2011; Moore et al., 2008); however, there is often at least one post-baseline assessment among these missing follow-ups.

Missing data, a common problem and a potential source of bias in research, involves information that is missing for some variable(s) and/or for some unit(s) of observation (Allison, 2001). In this thesis, missing data are defined as the absence of some value(s) on an outcome variable. Missing data also occur in the covariates but that is not the focus of this thesis. In practice, no matter how well designed and implemented, there will almost always be some missing data (Crutzen et al., 2013). Within RCTs, outcome data can be missing due to several reasons (Little & Rubin, 2002). For example, in a trial (Baerwald et al., 2010) with a 24% (191/810) dropout rate, reasons for the dropouts included lack of efficacy (n=63), adverse events (n=55), withdrawn consent (n=39), violation of eligibility criteria (n=18), loss to follow-up (n=7), and other unspecified reasons (n=9). A positive outcome, such as symptom relief, recovery, or cure, may also lead to discontinuation from a trial (National Research Council, 2010). Reasons for the dropouts are extremely important and should be collected, since they can be used to justify the assumptions of statistical analysis. Moore et al. (2008) examined participants' discontinuation in clinical trials based on 21 trial datasets in MSCs (osteoarthritis, rheumatoid arthritis, chronic low-back pain and ankylosing spondylitis) and reported that lack of efficacy or intolerable adverse events, or both, were the major reasons for discontinuation in those trials.

The validity and interpretability of findings from RCTs can be substantially reduced by missing data (Little & Rubin, 2002; Molenberghs & Kenward, 2007; European Medicines Agency, 2010; National Research Council, 2010; Fleming, 2011). In an RCT, as mentioned earlier, the advantages of randomization are jeopardized when the trial has missing data. To prevent selection bias in a clinical trial, it is important to adopt an ITT strategy, which requires all randomized patients to be included and analysed as randomized. However, the presence of missing data in a trial creates many challenges in

the selection of an ITT sample. The impact of missing data in a study is difficult to assess and is related to the question of what would hypothetically have been observed if no patient had withdrawn from the study. In general, there are three potential problems associated with the presence of missing data: loss of efficiency, bias in estimate of true parameters, and complication in data handling analysis (Horton & Lipsitz, 2001).

Loss of efficiency is an unavoidable consequence of missing data. Trials with missing data will be underpowered because fewer participants have completed than was originally planned; that is, the trial no longer has enough participants to demonstrate the same level of clinically important differences as statistically significant (Little & Rubin, 2002; Molenberghs & Kenward, 2007).

Another implication is that ignoring the presence of missing outcome data may lead to biased estimates, and thereby misleading inferences about treatment effects (Molenberghs & Kenward, 2007). A major concern is that being lost to follow-up could be related to a patient's responses to the treatment. Participants who do not complete a trial of a new treatment, for example, may be: those who improved the most, and do not see the necessity of continuing; those who improved the least, and see no reason to continue to comply with the treatment that is not working for them; or those who may have decided to discontinue owing to the occurrence of adverse effects. If the majority of dropouts are those who improved, then this will serve to make the interventions appear less effective than they actually are. Conversely, if most of the people dropped out because the new treatment was ineffective, this will, paradoxically, make the intervention look better, because many of the non-responders are no longer in that arm of the study. In view of the fact that missing data usually occur for reasons outside of the control of the investigators, and may be related to the outcome measurement of interest, the subsequent data analysis is extremely complicated.

As noted previously, no analysis should be treated as appropriate unless potential biases due to missing data are appropriatly addressed. To address this issue either imputation of values or modelling for missing data is generally required (European Medicines Agency, 2010), which rely heavily on untestable assumptions about the missing data – the wrong assumptions lead to biased estimates of treatment effect and standard errors. Since the potential impact of missing data depends primarily on missing data assumptions, it is important to investigate the processes (i.e. missing data mechanism) leading to missing data (Rubin, 1976; Little & Rubin, 2002).

2.3 Missing data: theoretical framework

To understand how best to deal with missing data, the first step is to determine the nature of the missing data and their possible implications for statistical inferences (National Research Council, 2010). The validity of any statistical analyses of incomplete data depends critically on causes of missing data. Since observed data cannot themselves explain definitely what might be the reasons for the missing data, it is necessary to make assumptions about the missing data mechanism. Therefore, statistical inferences on incomplete data rely on the subjective, untestable assumption about the distribution of missing data. Little and Rubin (1987; 2002) described a general missing data taxonomy, which includes a useful hierarchy of missing data mechanisms based on possible causal relationships between missing data and observed data in a study. Further discussions around this taxonomy for missingness in longitudinal data are available (Little, 1995; Schafer & Graham, 2002). A detailed review of this taxonomy is followed by an introduction of the repeated measures data structure.

Let $Y = (Y_1, Y_2, \dots, Y_T)$ refer to a vector of repeated measurements of an outcome variable Y on T occasions, and X as design variables that represent treatment indicators and baseline covariates. For simplicity, it is assumed that X is fully observed. To distinguish between observed and missing data, let $M = (M_1, M_2, \dots, M_T)$ denote the indicator of

whether Y is missing, where $M_{ij} = 1$ if observation at the j^{th} time for the i^{th} subject is missing.

2.3.1 Missing data patterns

In longitudinal studies, missed visits and/or study dropouts resulting in missing response data may occur. A missed visit occurs when a participant misses a clinic visit or fails to respond to a questionnaire meant for a particular follow-up visit during a follow-up schedule, whereas a dropout occurs when a participant discontinues from the study at any time during the study period and thus fails to provide outcome data thereafter. In trials with repeated follow-ups, participants who miss a study visit are often lost thereafter (National Research Council, 2010).

A dataset with a series of measurements on an outcome variable $Y = (Y_1, Y_2, \dots, Y_T)$ is said to have a monotone missing pattern when an event that a measurement Y_j is missing for an individual implies that all subsequent measurements $Y_k, k > j$, are missing for that individual. That is, under a monotone missing data pattern, the reason for missing data in a longitudinal trial is solely through study dropouts. Figure 2.1 shows a representation of monotone and non-monotone missing data patterns. A dataset with an arbitrary missing pattern is one with a monotone and non-monotone (i.e. intermittent) missing pattern; the missing data are due to both missed visits and study dropouts.

Monotone missing patterns

ID	Y_1	Y_2	Y_3	Y_4
1	o	o	o	o
2	o	o	o	m
3	o	o	m	m
4	o	m	m	m

Non-monotone missing patterns

ID	Y_1	Y_2	Y_3	Y_4
1	o	o	o	o
2	o	m	o	m
3	o	o	m	o
4	o	m	m	o

'o' – observed; 'm' – missing

Figure 2.1: Missing data patterns

2.3.2 Missing data mechanism

To understand the potential impact and how best to deal with missing data it is important to consider the process (i.e. mechanism) leading to the missingness. A general taxonomy for missing data, which is common in the statistical literature (Rubin, 1976; Little & Rubin, 1987; 2002; Little, 1995; Schafer & Graham, 2002; National Research Council, 2010), distinguishes between missing data that are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The classification is based on the dependence of missingness on observed and/or unobserved data. The missing data mechanism can be represented in terms of conditional distribution $[M \mid X, Y]$ for the missing data indicators given the values of the study variables that were intended to be collected.

Missing data are MCAR if missingness is independent of observed and unobserved data (i.e., missingness does not depend on values of the variables X and Y). That is, $[M \mid X, Y] = [M]$. This is the most desirable, but an unlikely scenario in trials with missing data. An example of such a scenario occurs when a participant discontinues a trial due to change of location during the course of the trial for reasons unrelated to the trial and/or disease of interest (e.g. job transfer). DeSouza et al. (2009) point out that missed visits are often not study-related, and thus the missing data are MCAR.

The second classification, which is more realistic than MCAR, is MAR. This mechanism requires that missingness is dependent on observed responses and/or covariates (X, Y_{obs}), but independent of unobserved responses (Y_{mis}). That is, $[M \mid X, Y] = [M \mid X, Y_{obs}]$. This type of missingness may be referred to as outcome-dependent MAR and/or covariate-dependent MAR, as the case may be (DeSouza et al., 2009). In longitudinal studies, MAR is plausible as dropouts are more likely to be related to previous responses (DeSouza et al., 2009). For example, if participants withdraw from a chronic pain trial once their pain intensity exceeds a certain threshold, then the missing data are MAR. The plausibility of

the MAR assumption can be improved by considering auxiliary variables that are predictive of whether the outcome variables are missing and predictive of the values of the missing variables (National Research Council, 2010).

MAR will fail to hold if missingness is dependent on unobserved data after accounting for available observed data. In that situation, the missingness is said to be MNAR. That is, $[M \mid X, Y] = [M \mid X, Y_{obs}, Y_{mis}]$. For example, if a participant feels better on his or her clinical condition after a visit and decides not to show up for the next scheduled visit, then the missing data are MNAR. Consequently, in longitudinal studies, future values of outcome variables for those who drop out cannot be reliably predicted based on data collected prior to dropping out if MNAR holds.

The implication of MCAR is that a missing data mechanism need not to be incorporated into an inference model, and a valid analysis is possible with observed data alone. For MAR, the missing data mechanism can be considered ignorable after including correlates of missingness in an inference model. For MNAR, on the other hand, the missing data mechanism is non-ignorable and needs to be incorporated into an analysis to make a valid inference. Importantly, a particular missing data mechanism is not in itself always ignorable or non-ignorable, depending on the statistical model. That is, if a statistical model fails to incorporate correlates of missingness, then a missing data mechanism cannot be considered as ignorable. Therefore, it is important to obtain additional variables that explain missingness and include these variables into the statistical model. As discussed here, missing data mechanisms play an important role in determining appropriate formal statistical analyses of data with missing values; however, it is difficult to distinguish the mechanisms in practice. In fact, on the basis of the observed data alone, it is impossible to identify the underlying missing data mechanism with certainty (Fielding et al., 2009).

2.3.3 Identification of the missing data mechanism

Few methods have been proposed to test for MCAR as a preliminary screening tool (Little, 1988; Diggle, 1989; Ridout, 1991; Fairclough, 2002). The purpose of these methods is not to explicitly detect violations of MAR, but violations of MCAR by identifying dependence on observed data. For example, Fairclough (2002) describes a logistic regression approach to confirm that a dropout process in repeated measurement data depends on observed data. A significant association between the dropouts and the observed data serves to rule out the possibility of MCAR.

Since there are valid analysis methods available for MAR data and not for MNAR data, the important consideration is the distinction between MAR and MNAR rather than between MCAR and MAR. To distinguish between MAR and MNAR, one must examine the relationship between missingness and unobserved data. Although it is impossible to determine the relationship empirically, a method has been proposed to evaluate the possibility of MNAR through a comparison of immediate responders who responded without any reminders and reminder responders who responded after sending reminders. By treating reminder responses as missing, Fielding et al. (2009) outlined an extension of Fairclough's logistic regression approach to determine whether the mechanism behind the reminder data are MNAR rather than MCAR or MAR. A significant difference in current scores between immediate and reminder responders after adjusting for the covariates that are predictors of reminder responses constitutes evidence of possible MNAR data. However, this evaluation excludes cases with actual missing responses.

2.4 Methods for handling incomplete continuous data

Several statistical procedures exist for handling missing data. These procedures can generally be divided into three broad categories: procedures based on listwise deletion; imputation-based procedures; and model-based procedures. Procedures based on

listwise deletion simply discard cases with missing values and analyse only those cases with complete data on all variables included in an analysis model. Analysis of covariance (ANCOVA), which is the most commonly used analysis method to estimate treatment effect in an RCT setting (Chapter 3), leads to listwise deletion of cases with missing data unless the missing values are imputed. In imputation-based procedures, missing values are replaced with particular values, which are determined by a specific procedure, in order to secure a complete dataset for analysis. As Little and Rubin (2002) explain, the purpose of imputation is to preserve important data characteristics, such as mean and variance, of the whole dataset but not to predict the true values of the missing data. From that perspective, a method that can replace missing values with multiple plausible values has been proposed (Rubin, 1978). Lastly, model-based procedures allow available data – not leading to listwise deletion – without imputation of missing values.

2.4.1 Description of candidate approaches to analysing longitudinal RCT data with missing values

2.4.1.1 ANCOVA without imputation of missing values

In an RCT, a transient benefit shown early in the trial might not persist through to the end of the trial and therefore it is necessary to demonstrate a sustained improvement at the primary endpoint. Although outcome data are commonly measured at more than one follow-up in trials, the aim of these trials is therefore usually limited to comparing the effect of two or more treatments at a specific time-point (Verbeke & Molenberghs, 2005). Unless it is important to know how study participants have reached the study endpoint, a simple comparison of the treatment groups at the primary endpoint is often recommended and adequate to demonstrate the treatment effect, if any (European Medicines Agency, 2006; Verbeke & Molenberghs, 2005). In case of continuous outcomes, a standard ANCOVA model, with baseline values of the outcome as the covariate, would be sufficient for such a comparison if outcome data are available on all participants (Van Breukelen, 2006;

Egbewale et al., 2014). Alternative methods such as analysis of variance and change-score analysis are not recommended because these analysis methods are subject to bias and are less precise than ANCOVA in relation to pretest-posttest correlation and the direction of baseline imbalance (Egbewale et al., 2014).

ANCOVA utilizes baseline and observed covariates as predictor variables in an analysis model, with the follow-up outcome at the primary endpoint as the outcome variable. For subject $i = 1, \ldots, m$ and repeated observations at visit $j = 0, \ldots, t$ (primary endpoint) per subject, the ANCOVA model is

$$Y_{it} = \beta_0 + \beta_1 Y_{i0} + \beta_2 X_i + \varepsilon_i,$$

where

- Y_{it}: outcome measurement at the primary endpoint for the i^{th} subject
- Y_{i0}: outcome measurement at baseline for the i^{th} subject
- β_0: intercept
- β_1: effect of baseline measurement (Y_{i0})
- β_2: effect size at the primary endpoint
- X_i: treatment group for subject i
- ε_i: assumed to be independently distributed from a univariate normal distribution

When employing ANCOVA, only subjects with complete observations on all the variables of interest are included in the model. This kind of approach is referred as complete-case analysis (CCA) in missing data literature. In this method, cases with missing data on any variable of interest are dropped from the analysis (i.e., listwise deletion of subjects with missing data). For example, Hewlett et al. (2011) investigated the effect of group cognitive-behavioural therapy compared to the control intervention on fatigue impact

among people with rheumatoid arthritis. The primary outcome – fatigue impact visual analogue scale – was measured at baseline, week 6, week 10 and week 18 (primary endpoint), and the study failed to measure the outcome on 33% (42/127) of participants at the primary endpoint. The primary analysis using ANCOVA removed those 42 participants irrespective of whether the outcome data were observed at earlier times. That is, the study analysed only a subset of participants. This generally does not provide a valid estimator of an ITT estimate (National Research Council, 2010). CCA requires the assumption that missing data are a random subset of the population of interest (Little & Rubin, 2002). Accordingly, if the missing data are MCAR, the sub-sample will be a random sample of the original sample, and the results of CCA will be unbiased but inefficient because of an inflated standard error if missingness is appreciable (Little & Rubin, 2002). However, if the missing data are MAR or MNAR, the analysis using CCA may not be valid as the reduced sample may no longer be representative of the population of interest, giving rise to biased estimates. In spite of these limitations, listwise deletion is still popular among researchers (Chapter 3) and is the default option with many statistical methods in major statistical software packages.

2.4.1.2 ANCOVA with single imputation of missing values

Single imputation methods replace each missing data point with a single value in order to produce a complete dataset to which standard statistical methods, such as ANCOVA, can be applied without discarding subjects with missing observations. That is, these methods treat imputed values as real values, and hence do not account for the uncertainty around the missing values. Therefore, the variance of estimates is likely to be too small, leading to underestimation of standard error. Further, these methods may produce biased estimates depending on how far the imputed value differed from the true value, which is unknown in real data. Hence, inferences based on the filled-in data can be distorted if the assumptions underlying the imputation method are invalid. Commonly, the imputation of

missing observations is based on the observed values. Several ad hoc strategies to perform single imputations – such as last observation carried forward (LOCF), mean imputation and regression imputation – are common in practice (Chapter 3).

Imputing by LOCF is a common single imputation method for repeated measures. In this method, a missing outcome value is replaced with the most recently available value of that outcome variable. LOCF therefore makes a strong assumption that there is no change in outcome for a participant after dropout. The rationale of this approach is that it is fairly conservative, as this approach likely underestimates the degree of change in an outcome over time (Streiner, 2008). However, this may not necessarily be the case in estimating a treatment effect (i.e. between-group difference in change) in trials, since imbalance between treatment groups in underestimation of degree of change in the outcome may overestimate the treatment effect. For example, if many participants who are expected to do worse over time discontinue a study treatment, or many of those who are expected to do well over time discontinue a control treatment, the benefit of the study treatment is more likely to be overestimated than the control with the LOCF method.

Some researchers contend that LOCF makes an MCAR assumption. For example, Mallinckrodt et al. (2008) state "*when assessing LOCF mean change via analysis of variance (ANOVA), the key assumptions are that missing data arise from an MCAR mechanism and that for subjects with missing endpoint observations, their responses at the endpoint would have been the same as their last observed values*". However, the assumption of no change in outcome after drop out may not be valid under MCAR or MAR (National Research Council, 2010). As mentioned earlier, the MAR assumption is that the predictive distribution of an outcome variable at time t (Y_t) conditional on design variable X and observed data $(Y_1, \dots, Y_{t-1})$ is the same for both observed and missing Y_t. However, LOCF assumes that the predicted value of missing Y_t is Y_{t-1} with probability one and the

assumption may not hold with observed Y_t (i.e. the probability may not be one for the observed data). That is, in general LOCF makes an MNAR assumption.

2.4.1.3 ANCOVA with multiple imputation

Multiple imputation (MI) is designed to substitute each missing value with a range of plausible values based on observed information in order to reflect the uncertainty associated with imputation of missing values (Rubin, 1978; 1987; 1996; Schafer, 1997; Little & Rubin, 2002; Molenberghs & Kenward, 2007; Carpenter & Kenward, 2013). For example, in a study with two observations (age and pain intensity) on each participant, suppose that age is completely observed and pain intensity is incomplete. The basic idea is to use the association between age and pain intensity from the complete data to fill the missing pain intensity score with multiple plausible values. These multiple values are a random draw from the posterior predictive distribution of missing values based on a statistical model explaining the association between the variables. The imputation is based on an implicit and untestable assumption that the association between age and pain intensity is the same for those participants who provided complete data and those who do not. That is, MI requires the MAR assumption.

The MI procedure involves three stages (Rubin, 1987): imputation, analysis and pooling. Initially, it is required to specify a statistical model (referred to as an imputation model) to explain the relationship between observed data and missing data. The posterior distribution for the estimated parameters of the model is used to simulate the parameters of the posterior predictive distribution of the missing data from which *m* predicted values are drawn. This creates *m* completed datasets. In the second stage, one fits a statistical model (referred to as analysis model), e.g. ANCOVA model, to each completed dataset, and generates parameter estimates $\hat{\beta}^1, \hat{\beta}^2, \ldots, \hat{\beta}^m$ and associated variance. Finally, these estimates and variances are combined into a summary inference about β, using Rubin's

rule (1987). The MI estimate is the average of the estimates from the m datasets, and the variance of the estimate is

$$W + \left(\frac{m+1}{m}\right) B,$$

where W is the average of the variances from the m datasets and B is the between-sample variance of the estimates over the m datasets.

Methods for MI of multivariate missing data include three broad approaches (Molenberghs & Kenward, 2007): (i) sequential MI for data with monotone missingness; (ii) joint modelling (JM) approach, which assumes that all variables in the imputation model jointly follow a multivariate normal distribution; and (iii) full conditional specification (FCS) approach (which is referred to as multiple imputation by chained equation [MICE]), which does not rely on the multivariate normality assumption. If the missing data pattern is monotone, regression imputation can be performed sequentially, starting with the variable having least missing values; this approach uses noniterative techniques for simulating from the posterior predictive distribution of missing data. The imputation method based on a multivariate normal regression uses an iterative Markov chain Monte Carlo (MCMC) technique to simulate from the posterior predictive distribution of missing data. The MICE method uses a Gibbs-like algorithm to obtain imputed values. If the missingness mechanism is non-monotone, both JM and FCS approaches are generally preferred and provide similar results in a standard regression analysis involving a mixture of continuous and categorical variables (Lee & Carlin, 2010).

2.4.1.3.1 Selecting variables for an imputation model

As mentioned earlier, MI requires two statistical models: an imputation model, which is used to impute missing values and a substantive analysis model, which is used to analyse the imputed data. Choice of imputation model for the MI can have a pronounced effect on

the outcome of the data analysis (Spratt et al., 2010). Ideally, an imputation model should include all variables that are in the substantive analysis model and should reflect the structure of the subsequent analysis (Kenward & Carpenter, 2007; Sterne et al., 2009; Carpenter & Kenward, 2013) – which is referred to as a restrictive modelling strategy (Collins et al., 2001). Further, it is possible to incorporate auxiliary variables that are not part of the analysis model into the imputation model in order to make MAR more plausible and, therefore, to increase efficiency and reduce bias (Collins et al., 2001; Spratt et al., 2010) – which is referred to as an inclusive modelling strategy (Collins et al., 2001). White et al. (2011b) pointed out that one should include in an imputation model all variables that predict the incomplete variables in an analysis model and/or predict whether the observations on the incomplete variables are missing.

2.4.1.3.2 Selecting the number of imputations

The amount of missing data should be considered when deciding the number of imputations, say *m*, in the MI method. Rubin (1987) previously demonstrated the relative efficiency of a finite-*m* estimator as $\frac{V(\bar{\theta}_\infty)}{V(\bar{\theta}_m)} = \left(1 + \frac{\gamma}{m}\right)^{-1}$, where γ is the fraction of missing information for an outcome measure to be analysed. Using this formula, table 2.1 shows the relative efficiencies with different m and fractions of missing information.

Table 2.1: Percent efficiency of MI estimation

m	Rate of missing data (γ)				
	0.1	0.3	0.5	0.7	0.9
1	91	77	67	59	53
3	97	91	86	81	77
5	99	94	91	88	85
10	99	97	95	93	92
∞	100	100	100	100	100

Following Rubin's calculation of relative efficiency, many researchers have advocated a small number of imputations as adequate to yield excellent results. For example, Schafer and Olsen (1998) suggested only 2–5 imputations, and Schafer (1999) further emphasised that no more than ten imputations are usually required. However, these suggestions have been recently critiqued. Graham et al. (2007) and Spratt et al. (2010) observed that, with the small number of imputations, variability due to the imputation procedure was substantial enough to affect inferences, and recommend that many more imputations should be performed than previously considered adequate. In favour of this recommendation, White et al. (2011b) proposed a rule of thumb, which states that the number of imputations should at least be equal to the percentage of incomplete cases.

2.4.1.4 Model-based methods: direct maximum likelihood estimation

Maximum likelihood (ML) refers to a method of estimating the parameters of a statistical model. ML estimates, which maximize the likelihood function of sample data, are asymptotically unbiased if the model has been specified correctly (Little & Rubin, 2002). When data is incomplete, direct likelihood based methods (also referred to as full-information ML [FIML] methods) can use all available data, instead of deleting observations with missing values, for analysis without explicit imputation of missing data, and assume that the missing data mechanism is either MCAR or MAR (Little & Rubin, 2002). The joint likelihood for the observed data y_{obs} and the missing data indicator m is:

$$
\begin{aligned}
L(\theta, \varphi \mid y_{obs}, m) &= \int f(y_{obs}, y_{mis}, m;\ \theta, \varphi)\, dy_{mis} \\
&= \int f(y_{obs}, y_{mis};\ \theta)\, f(m \mid y_{obs}, y_{mis};\ \varphi)\, dy_{mis} \\
&= \int f(y_{obs}, y_{mis};\ \theta)\, f(m \mid y_{obs};\ \varphi)\, dy_{mis} \text{ if MAR} \\
&= f(y_{obs};\ \theta)\, f(m \mid y_{obs};\ \varphi)
\end{aligned}
$$

Thus, if the parameters θ and φ are distinct, then θ can be estimated by maximizing the observed data likelihood $f(y_{obs}; \theta)$ alone, independent of the model for m. Therefore, when the missingness mechanism is MCAR or MAR, specification of a missingness model is unnecessary and inferences are based on the likelihood function given the observed data only (DeSouza et al., 2009). In addition to the MAR assumption, FIML methods require a large sample size and need to meet the multivariate normality assumption for the variables used in a model (Little & Rubin, 2002).

2.4.1.4.1 Mixed-effects models (Random-effects models)

For longitudinal data, mixed-effects models provide a parsimonious way to specify a multivariate distribution. Linear mixed models are an extension of linear regression models allowing for inclusion of random-effects to account for within-subject dependency in the longitudinal measurements (Laird & Ware, 1982). Specification of mixed-effects models requires: (i) a model for the mean structure of the longitudinal data, which usually depends on covariates, design matrix for time, treatment group and patient specific random-effects; (ii) an assumption on the distribution of the random-effects; and (iii) specification of an additional correlation matrix in the longitudinal measurements (Wong et al., 2011). The model can be specified as

$$\mathrm{Y}_i = X_i\beta + Z_i\upsilon_i + \varepsilon_i,$$

where

- Y_i is the n_i x 1 vector of responses for subject i, and n_i is the number of measurements for subject i ($i = 1 \text{ to } m$)
- X_i is a known n_i x p covariate matrix of i^{th} subject for fixed effects β
- Z_i is a known n_i x q covariate matrix of i^{th} subject for random effects υ_i
- β is a p x 1 vector of unknown population parameters

- v_i is a q x 1 vector of unknown subject effects (random-effects) distributed as $N\ (0, D)$
- ε_i is a n_i x 1 vector of random residuals distributed independently as $N\ (0, \Sigma_i)$
- v_i and ε_i are independent

In matrix notation, $\boldsymbol{G} = \begin{pmatrix} D & & \\ & \ddots & \\ & & D \end{pmatrix}$ and $\boldsymbol{R} = \begin{pmatrix} \Sigma_1 & & 0 \\ & \ddots & \\ 0 & & \Sigma_m \end{pmatrix}$ represent the variance-covariance matrices of v and ε respectively. Therefore, the variance-covariance matrix for the vector of outcomes for all subject visits $\boldsymbol{Y}$ is specified as:

$$\boldsymbol{V} = \mathrm{Cov}\,(\boldsymbol{Y}) = Z\boldsymbol{G}Z^T + \boldsymbol{R}$$

2.4.1.4.2 Mixed-effects model for repeated measures using categorical time effects (MMRM)

Likelihood-based, mixed-effects models offer a general framework to extend the standard ANCOVA for repeated measures data in a clinical trial to provide the direct estimates and statistical test for treatment group differences at the primary endpoint, which is typically of direct interest for regulatory decision-making, while incorporating available data for participants who dropped out early (Mallinckrodt et al., 2001a). In many longitudinal RCT settings, the repeated measures are balanced in the sense that outcomes are assessed at the same time interval over a limited number of visits for all participants. This allows a "saturated" mixed-effects model to be specified by including a full treatment group by measurement time interaction for outcome means combined with an unstructured within-subject error variance-covariance matrix[1] (Beunckens et al., 2005). The model is often

[1]An unstructured within-subject error variance-covariance matrix is recommended unless the underlying ("true") structure is known (Mallinckrodt et al., 2004)

referred to as MMRM (Mallinckrodt et al., 2001a; 2003). MMRM is a particular parameterization of the mixed linear model, often with no random-effects,[2] and with time of assessments considered as a factor variable and group-by-time effect as an unstructured interaction effect, instead of considering the group-by-time effect as the slope difference of the treatment groups over time (Mallinckrodt et al., 2001a; 2003). That is an MMRM model does not make any assumptions about the shape of the response profile over time. Since there is no random-effect specified (i.e. Z and $\boldsymbol{G}$ are zero), the variance-covariance matrix for the vector of outcomes for all subject visits $\boldsymbol{Y}$ becomes:

$$\boldsymbol{V} = \mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{R}$$

2.4.1.4.3 Strategies to model baseline responses

Various strategies on how to handle baseline responses in an MMRM analysis are discussed in the literature (Liang & Zeger, 2000; Fitzmaurice et al., 2004; Liu et al., 2009; Kenward et al., 2010; Dinh & Yang, 2011). Among those, two strategies are generally recommended in RCT settings, where detecting treatment effect at a particular time-point is the major consideration (Fitzmaurice et al., 2004; Dinh & Yang, 2011). They are:

i. Consider the baseline responses as a covariate in the analysis of follow-up responses, allowing different regression slopes by specifying a 'baseline-by-time' interaction term into the MMRM model.

ii. Retain the baseline responses as part of the outcome vector and assume mean responses at baseline are equal between the groups in the MMRM model. This approach sometimes referred as constrained longitudinal data analysis (cLDA).

---

[2]The specification of the random part '|| *id:*' with Stata mixed procedure statement is not to introduce random-effects at the subject level, but to group the repeated measurements as a "blocking factor", where data from separate subjects are uncorrelated and data within each are correlated; where variable *id* is a unique identifier of subjects.

The additional restrictions in both strategies provide an adjustment for the observed baseline difference in estimating the treatment effects. In practice, it is quite common that some participants may miss either baseline assessment or entire follow-up assessments. The systematic review of trials in chapter 3 reports that there are instances of randomized participants who discontinued the trial and failed to provide outcome responses after baseline visits; thus these trials failed to perform a true ITT analysis. The model with baseline as an outcome (i.e. cLDA) provides a framework for including all randomized participants in the analysis who have baseline or follow-up assessments. Whereas the usual MMRM model with baseline as a covariate includes only those individuals who have baseline response and at least one follow-up assessment. These methods give identical point estimates and very similar SEs when the baseline is complete (Liang & Zeger, 2000; Kenward et al., 2010; Dinh & Yang, 2011). In contrast, Fitzmaurice et al. (2004) and Liu et al. (2009) favour treating baseline values as an outcome, and Liu et al. (2009) found that retaining baseline as a covariate could result in slightly greater loss of efficiency compared to the other approach. However, Kenward et al. (2010) argue that Liu et al.'s (2009) conclusion is flawed because of failure to use restricted ML instead of ML estimation and a correction for finite sample bias (for example, the Kenward–Roger adjustment in SAS *proc mixed*). They further commented that the loss of efficiency was accompanied by a gain in confidence interval (CI) coverage.

2.4.2 Choice of approach to handling missing data: existing literature on comparisons of the approaches

Missing values in a clinical trial data lead to concern and confusion in identifying the full dataset according to an ITT approach, making data analyses more complex and challenging. Though a number of approaches are discussed in the literature to deal with missing values in a clinical trial none of them can be regarded as a universal approach because each trial has its own design and measurement characteristics. As discussed

earlier in this chapter, the above discussed approaches to deal with missing outcome data were developed under certain missing data assumptions. That is, methods leading to listwise deletion assume MCAR; LOCF assumes outcome remains constant after dropout of a participant; MI and MMRM assume MAR. Identification of the underlying missing data mechanism is important in order to carry out appropriate formal analyses of data with missing values. However, it is impossible to identify this mechanism with certainty based on the observed data alone (Fielding et al., 2009). Therefore, to aid the selection of an appropriate approach that best deals with missing data, the relative performance of missing data techniques needs to be considered under various clinical trial scenarios. The guideline for confirmatory clinical trials (European Medicines Agency, 2010) specifies that the primary analysis of clinical trial datasets can only be accepted if the analysis is considered to be reasonably free from biases that favour the experimental treatment, and if it can be verified that the variability of the estimated treatment effect is not underestimated to an important extent.

Collins et al. (2001) claim that both MI and direct likelihood based approaches tend to yield very similar estimates when both analyses are implemented in a similar manner (i.e. imputation model is similar to likelihood-based model). However, little research has been done in comparing the performance of MMRM and MI in controlling the type 1 error rate and statistical power in hypothesis testing of treatment effect. Barnes et al. (2008) performed a simulation study to evaluate the type 1 error rate in baseline observation carried forward (BOCF), LOCF, MMRM and MI where no differences existed between treatment groups in mean change to endpoint. Their study was designed to mimic several clinical trial data characteristics with high differential dropout rates between study groups (high dropout rates in a placebo-controlled group [25% vs 33%; 25% vs 40%] and vice versa) under an MNAR mechanism. The study found that both MMRM and MI outperformed BOCF and LOCF in most scenarios. However, the study reported a slightly larger average standard error from MI compared with MMRM, thus resulting in a wider CI

with MI. A wider CI controls type 1 error conservatively under a null hypothesis, and fails to provide adequate statistical power under an alternative hypothesis. Barnes et al. (2008) argue that the larger standard error may be associated with the low number of imputations ($m = 5$) they used with MI.

Siddiqui (2011) reported a similar finding that supports MMRM as a better choice against MI, even under MAR dependent on observed outcome data. Siddiqui (2011) performed a simulation study to assess the relative performances of MMRM, MI (JM approach), and MI (sequential approach) in controlling type 1 error and statistical power. Their study considered two correlation matrices (strong and moderate) along with high differential dropout rates (placebo: 30% vs new treatment: 40%) between study groups to simulate datasets under a MAR mechanism dependent on observed outcome data. The study used a relatively high number of imputations ($m = 10$) compared to that in Barnes et al. (2008). In this simulation study, both MI procedures yielded very similar results in all scenarios. When the null hypothesis was true, MMRM and MI analyses estimated the null effect, but MI was too conservative in controlling type 1 error. On the other hand, when there was a true difference between study groups, MI analysis underestimated the true difference, and it produced a larger standard error of the estimate. The combined effect was substantially lower statistical power with MI analysis in comparison to the corresponding power with MMRM analysis. The bias in the estimate raises concern over the simulation study implementation since both MMRM and MI were developed under an MAR mechanism, and these analyses are expected to produce unbiased estimates. Additionally, the substantial difference in statistical power between the MMRM and MI analyses in this study may be partially explained by the lower number of imputations, against the recent suggestion by White (2011b), and estimation using ANCOVA instead of MMRM after MI. However, this explanation needs to be verified through an extensive simulation study. Peters et al. (2012) did a simulation study to assess the added value of MI of missing outcome values in longitudinal RCT datasets (overall dropout rate ranged 10%–60%)

analysed with linear mixed-effects (LME) models, and found no additional benefit for doing so. Although the study questioned the advantage of MI in analysing missing repeated outcome measurements, the comparison of the findings to Siddiqui (2011) is limited due to several factors. Importantly, the dropouts were generated under an MAR assumption dependent on a fixed covariate (age at recruitment). Additionally, the study was a comparison of LME analyses with and without MI, and the estimand was the difference in rate of change (slope) between study groups instead of between-group difference in treatment effect at an endpoint.

Like Barnes et al. (2008), Olsen et al. (2012) performed a simulation study under MNAR scenarios to compare CCA, LOCF, MMRM and MI (JM and sequential approaches; $m = 10$) in terms of bias, type 1 error, and power. Deletion of outcome values at follow-up visits was implemented – in a similar manner reported by Baron et al. (2008) – dependent on the rate of change from the previous visit to the current visit. This deletion criterion led to two scenarios: (i) nearly the same dropout rate in both groups (31% with true null hypothesis and 34% with true alternative hypothesis); (ii) slightly differential dropout rate between the groups (placebo: 36% vs new treatment: 31% with true null hypothesis; 34% vs 29% with true alternative hypothesis). Interestingly, all but the MI (JM) approach retained the type 1 error rate within an acceptable range[3] in nearly all scenarios considered in the study. Slightly larger type 1 error rate with CCA had been reported under the differential dropout scenario. Importantly, MI (JM) was too conservative in controlling type 1 error. With an alternative hypothesis, their study found notable differences in bias and power between equal and differential dropout scenarios; bias and loss of power were minimal with CCA, MMRM and MI (sequential) under the equal dropout scenario. In all but the MI (JM) method, the observed power was higher than 80%

[3] Burton et al. (2006): acceptable range 3.6%–6.4% was determined based on 1000 receptions in the simulation study.

in all scenarios with the high dropout rate against the observed power of 96% with the original dataset having no missing data; for MMRM, it was higher than 90% with the high dropout. Olsen et al. (2012) argue that the poor performance of MI (JM) was due to the exclusion of a variable indicating randomized treatment group from the imputation model. Lyass (2010) recommends the inclusion of randomized treatment group as a covariate in the imputation model to maintain type 1 error and power of the subsequent statistical test comparing treatment effect.

The simulation study by Olsen et al. (2012) reported a higher power along with a slightly lower CI coverage when MMRM was used instead of ANCOVA, irrespective of whether MI (sequential) was being used or not. Their study also reported similar estimates of treatment effect, type I error rate and statistical power from MMRM or ANCOVA, irrespective of MI (sequential) being used or not. That is, this study did not find any marked benefit of employing MI of missing repeated measurements in an RCT dataset prior to analysing with MMRM or ANCOVA. However, the results that suggested no substantial differences in estimate of bias, power and CI coverage between CCA and MI ANCOVA under MNAR scenarios of a high dropout rate is a matter of concern. In summary, these findings favoured an MMRM analysis of available data without any kind of imputations under the scenarios considered in their study. However, this summary contrasts markedly with the findings of a simulation study by Baron et al. (2008), which had been implemented with similar dropout criteria but lower differential dropout between groups (placebo: 20% vs new treatment: 11%). Baron et al. (2008) supported MI (JM; m=5) against other approaches (listwise deletion, LOCF and LME) used in this study in respect of good control over type 1 error rate, power and bias. Further, in this simulation study, MI (JM) substantially outperformed both CCA and LME. Since their study used ANOVA and LME modelling as the analysis model instead of ANCOVA and MMRM, and it failed to retain the nominal CI coverage of 95% for the estimates from the original data having no missing values even with 1000 replications in the simulation, a direct

comparison with Olsen et al. (2012) may not be ideal. Additionally, a higher number of follow-up visits (6 visits), in comparison with Baron et al. (2008) (3 visits), may contribute to favouring longitudinal data modelling in Olsen et al. (2012), but the lower dropout rate in Baron et al. (2008) cannot be ignored.

Another important finding from Olsen et al.'s (2012) simulation study was that CCA, MMRM, and MI (sequential) could estimate the true treatment effect in the original data without missing values when the dropout rate was equal between the groups. With equal dropout rate between the groups, these methods could also retain the CI coverage in the acceptable range[4] and the loss of power in a minimal level. Recently, Bell et al. (2013) demonstrated that equal dropout does not guarantee unbiased estimates of treatment effect through a simple simulation study, which compared CCA, LOCF and MMRM under different missing data mechanisms with a high overall dropout rate.

In the above simulation studies, by ignoring the limitations discussed, either MMRM or MI was the superior approach to deal with missing outcome data in longitudinal clinical trials in comparison with CCA and single imputation approaches. In addition, several simulation studies have addressed the impact of missing data and compared methods of handling missing outcome data in longitudinal clinical trials (Mallinckrodt et al., 2001b; 2001a; 2004; Lane, 2008; Siddiqui et al., 2009; DeSouza et al., 2009). The first three studies (Mallinckrodt et al., 2001b; 2001a; 2004) had been performed to compare MMRM (but not adjusted for baseline observations) against LOCF ANOVA, and aimed to understand how robust these approaches are to violations of the MAR assumption. Mallinckrodt et al. (2001b) evaluated these approaches in a range of deletion criteria, and found that the MMRM provided adequate control of type 1 error rate even in the most extreme scenarios

[4] Burton et al. (2006): acceptable range 93.6%–96.4% was determined based on 1000 receptions in the simulation study.

(i.e., scenarios with high dropout rates (>30%) and a strong MNAR mechanism). In contrast, type 1 error rates from LOCF ANOVA were inflated severely in most scenarios due to an increased bias – a similar observation was later made by Barnes et al. (2008). In a simulation study with high differential dropout rates between treatment groups (placebo: 60% versus new treatment: 30%), Mallinckrodt et al. (2001a) observed that the average estimates of treatment effect from the MMRM model did not substantially deviate from the true value, and the standard error and CI accurately reflected the true uncertainty of the estimates. On the other hand, LOCF ANOVA produced biased estimates – markedly underestimating the difference in treatment effect when both treatment groups in their simulation study were effective and overestimating the difference when one of the treatments was ineffective. In another simulation study that evaluated the effect of within-subject error correlation structure, Mallinckrodt et al. (2004) reported that specifying an unstructured matrix for use in MMRM, regardless of the true correlation structure, yielded superior control over type 1 error, bias and CI coverage than LOCF in every scenario. The previous work by Lane (2008) and Siddiqui et al. (2009) supports MMRM against LOCF ANCOVA in respect of good control over bias in the estimate of treatment effect and CI coverage in all missing data mechanisms. However, unlike the studies by Mallinckrodt et al. (2001b; 2001a; 2004), they reported that both methods were affected by serious misinterpretation due to biased estimates of treatment effect and coverage in MNAR data. Unlike the other studies mentioned in this paragraph, DeSouza et al. (2009) compared the performance of CCA, LOCF and MMRM through simulated datasets with relatively smaller overall dropout rates (10%, 15% and 30%). Their study evaluated a number of dropout patterns under an MAR dropout mechanism, and reported that MMRM performed consistently well – in terms of controlling bias – compared to the other methods across the dropout rates and dropout patterns; however, the loss of power was substantial in all these methods.

2.4.3 Sample size calculation in anticipation of dropouts

Sample size calculations are essential in the proper design of an RCT. Particularly important in relation to missing data is how to account for loss of power due to dropouts in hypothesis tests or CIs; however, this issue has not been studied extensively (Little et al., 2012). The most intuitive strategy would be to multiply by some factor the number of subjects determined under an assumption of no dropout. For example, if an anticipated dropout rate is d ($0 < d < 1$), then the multiplication factor would be $1 + d/(1 - d)$. The inflation of sample size is based on an assumption that the extent of loss in nominal power is proportional to the amount of missing data. However, this assumption is not necessarily true, and it is important to take into account the effect of any imputation methods that will be used in the analysis.

A few of the previous simulation studies (Mallinckrodt et al., 2001a; Baron et al., 2008; Lane, 2008; DeSouza et al., 2009; Siddiqui, 2011; Olsen et al., 2012) compared various missing data handling approaches in terms of statistical power, and found that all the approaches influenced the expected power irrespective of biased or unbiased estimates of treatment effects from these approaches. As evident from the previous simulation studies, the achieved power of a study given an analysis method is influenced by the amount and direction of bias – the bias in estimate leads to artificial inflation (due to substantial overestimation of treatment effect) or reduction of power. Additionally, even methods that provide unbiased estimates of treatment effect may not protect against loss of power. However, it is unclear whether the loss of power is a matter of concern with an inflated sample size, which is calculated using the naïve approach to account for the loss of power due to attrition. These simulation studies used an unrealistic sample size (due to random selection of sample size) that yielded substantially higher or lower power in the absence/presence of missing data than the routinely used desired power of 80% or 90% in real practice. As was pointed out in the previous section, three studies – Mallinckrodt et

al. (2001a); Baron et al. (2008); Olsen et al. (2012) – limited their exploration to an MNAR mechanism. Siddiqui (2011) did the exploration under an MAR mechanism but reported biased estimates of treatment effect with MI. Lane (2008) and DeSouza et al. (2009) did not consider MI in their simulation studies. Further, none of these studies explored the impact of different sample size or different level of dropout rate to understand its effects on statistical power. Therefore, the findings from these simulation studies might have limited generalizability in real practice. It remains unclear how the extent and distribution of missing data actually influences the power in longitudinal clinical trials, and whether the naïve approach is helpful to attain the desired power in the missing data scenarios.

Lu at al. (2008; 2009) have proposed a modified multiplication factor to increase the number of subjects determined under an assumption of no dropout where the estimation of treatment effect at an endpoint is based on an MMRM model. The approach requires complete specification of correlation matrix and dropout rate by treatment group over time. However, in practice there is often only limited data at the design stage of RCTs. Therefore, the use of this approach is limited by the extensive computational requirements and pre-specification of more information on parameters.

2.5 Guidance on prevention and handling of missing data in RCTs

Several sets of guidelines and recommendations on prevention and handling of missing data in RCTs have been published in order to promote best practice, and to maintain standards on study design, study conduct, data analysis and reporting (Mallinckrodt et al., 2008; European Medicines Agency, 2010; National Research Council, 2010; Little et al., 2012; Dziura et al., 2013; Li et al., 2014). Li et al. (2014) conducted a systematic review on regulatory guidelines containing recommendations relevant to the prevention and handling of missing data in clinical studies, including RCTs. The review concludes with three important points for consideration. First, the single best approach is to take appropriate measures at the design stage to prevent the occurrence of missing data, as

there is no single method to solve the problem of missing data. Second, trialists should use valid statistical methods that properly reflect all the sources of uncertainty in respect of the missing data. Third, details of missing data should be reported thoroughly and transparently to allow readers to judge the validity of the findings.

With particular reference to RCTs, two documents – “The prevention and treatment of missing data in clinical trials” by U.S. National Research Council (NRC) panel (2010) and “Guideline on missing data in confirmatory clinical trials” by European Medicines Agency (EMA) (2010) – provide extensive coverage of the missing data problem. The NRC report provides 18 recommendations to address missing data in clinical trials through careful trial design, conduct and analysis. Recommendations 9–15 specifically provide important direction towards the analysis of missing data. The EMA guideline provides advice on how missing data in confirmatory clinical trials should be addressed and reported, and on regulatory acceptability of these approaches. Unlike the NRC report, this guideline covers only general recommendations on acceptable frameworks for handling missing data in a regulatory setting. This guideline discourages the use of CCA as a primary analysis. However, the guideline does not completely agree with the NRC report on the use of LOCF or BOCF; these approaches can be acceptable if they can be shown to be conservative. This guideline also supports the use of either MI or mixed models as the primary analysis, and concludes by indicating the importance of sensitivity analyses that make different assumptions to assess the robustness of trial findings.

2.6 Rationale of the thesis

As detailed in section 2.4.2, there have only been a small number of simulation studies evaluating the performance of statistical methods in the presence of missing data in respect of treatment effect in RCTs. Of these, some compared the performance of MMRM and MI in respect of bias, CI coverage and statistical power, while the remaining studies reported comparisons between MMRM and LOCF-based analysis or CCA. None of the

previous simulation studies favoured either CCA or LOCF under an MAR or MNAR mechanism; however, the findings on MMRM and MI substantially varied across these studies. The studies largely had a narrow real-life clinical scope and limited applicability. Hence, due to the lack of wider generalizability of these previous studies, a comprehensive simulation study is planned to examine the relative performance of missing data handling strategies on a number of possible and credible clinical trial scenarios in order to provide a broader and practically more accessible picture of the impact of missing outcome data on the estimation of treatment effect in an RCT.

The efficiency and accuracy of estimates from statistical methods depend on how close mechanisms truly generating the missing data are to the underlying statistical assumptions of the methods used. In practice, the missing data mechanisms are not strictly identifiable from incomplete data, so the desired "fit" in terms of assessing whether a method properly aligns to a mechanism of missingness is difficult to ascertain. As pointed out in section 2.3.3, although some methods have been proposed for the identification of missing data mechanism, their purpose is generally to detect violations of the MCAR assumption by identifying dependence of missingness on observed data, but not to confirm either MAR or MNAR assumption. Since it is not possible to guarantee a particular missing data mechanism on an incomplete dataset, it is important to assess the sensitivity of primary analysis results to departure from a missing data mechanism that was assumed in the analysis. Importantly, MNAR-based analyses require stringent assumptions about missing data and are rarely reported in practice (details will be discussed in chapter 3). The NRC report on the prevention and treatment of missing data in clinical trials (National Research Council, 2010) highlights the need for sensitivity analyses to confirm the primary analysis findings; however, the report also acknowledges the lack of guidelines on the selection of sensitivity analyses and interpretation of their findings, and lack of software packages to implement such analyses. Due to such difficulties associated with the sensitivity analyses, in this thesis I propose a simple

approach using the responses obtained after a number of failed attempts to verify the ignorability of the missing data that is assumed by the primary analysis and hence the unbiasedness of the estimate of treatment effect. The results from the proposed approach will be a “sign-post” to decide whether it is important or not to proceed to carry out sensitivity analyses that assess the robustness of the original estimates to deviation from the ignorable assumption relating to the mechanism for missingness.

2.7 Conclusion

RCTs are the gold standard for evaluating the direct causal relationship between intervention and outcome, as randomization equalizes known and unknown characteristics between intervention groups. Since the presence of missing data may affect the balance achieved through randomization, RCTs may fail to recognize the underlying causal relationship. Further, an analysis of incomplete data may be less efficient and leads to biased estimation of treatment effect. Despite efforts to minimize missing data, the problem will inevitably still occur in RCTs. Any analysis of incomplete data requires unverifiable assumptions about the nature of the missing data, and validity of inferences from these analyses depends on the correctness of these assumptions. Since, it is not possible to exactly verify the correctness of the missing data assumption based on observed data alone, it is important to assess the robustness on the inferences to a variety of scenarios in RCT settings.

In this chapter, I have discussed four methods to deal with incomplete longitudinal continuous outcome data: two frequently used ones – standard ANCOVA and LOCF ANCOVA – and two frequently recommended ones – MI ANCOVA and MMRM. As discussed earlier, a few simulation studies have been performed in respect of a limited number of scenarios – especially those featuring a high overall dropout rate – to evaluate the performance of these methods. Due to the limitations of these studies and their contrasting findings discussed earlier, a comprehensive simulation study is required to

evaluate these methods. The need for further research has been advocated in the NRC report (National Research Council, 2010). The areas in need of further research include: (i) the effect of missing data on the power of clinical trials, (ii) how to set useful target rates and acceptable rates of missing data in clinical trials, and (iii) the robustness of missing data methods to violations of its assumptions.

Chapter 3: Systematic review

3.1 Introduction

In this chapter, an attempt is made to review current practices in analyses of randomized clinical trial (RCT) data in the presence of protocol violations (e.g. inclusion of ineligible patients, treatment crossover) and missing data. Here, the review focuses on RCTs reported in five leading medical journals that mainly publish studies of musculoskeletal conditions (MSCs). Trials on MSCs differ from those in other areas, due to several characteristics that may influence the trials and lead to missing outcome data in many situations. The review seeks to cover current issues and practices related to statistical analyses of RCT data centred on handling of missing outcome data, with specific reference to trials in MSCs. The objectives of this review are listed in detail in section 3.3.

3.2 Background

Intention-to-treat (ITT) analysis is the preferred method of analysis for RCTs with a superiority design. The ITT principle states that an analysis should be performed by including all study participants in the groups to which they were randomized, regardless of any departures from the original assigned group (Chan et al., 2013). This principle helps to preserve the benefits of randomization, which is intended to ensure that differences in outcome observed between the groups are solely the result of the treatment (Montori & Guyatt, 2001; Heritier et al., 2003), and to reduce the risk of selection bias (Altman, 2009; Fleming, 2011). In an ideal setting, all subjects enrolled in an RCT would follow the study protocol and complete their allocated treatment as detailed therein, and thus contribute data that were complete in all respects (Lewis & Machin, 1993). However, this is rarely achieved in practice – particularly under pragmatic trial conditions. Moreover, to provide an unbiased estimate of treatment effect, randomization alone is not sufficient and it is also important to obtain complete data on all randomized subjects (Lachin, 2000). Some

authors, however, describe an analysis as ITT without regard to this requirement to include data for all randomized participants in the analysis (Higgins & Altman, 2011). In this review, an approach that deviates from a full ITT analysis in this way – by retaining treatment group membership as per random allocation but excluding participants with no follow-up data – was referred as a partial ITT analysis. The term "modified ITT" has frequently been used to describe this approach (Higgins & Altman, 2011), but this term has been criticized for being ambiguous and lacking clarity regarding the exclusion of data (Ioannidis et al., 2004; Abraha & Montedori, 2010).

Due to a perceived misuse of the term "ITT" (Ioannidis et al., 2004; Abraha & Montedori, 2010; Schulz et al., 2010), item 16 in the 2010 CONSORT statement was updated to include a more explicit request for group-wise details on the number of participants included in each analysis and whether the analysis was by randomized groups (Schulz et al., 2010). Non-ITT analyses such as an 'as-treated' (AT) analysis, which groups participants according to treatment received rather than according to randomization, and a 'per-protocol' (PP) analysis, which omits participants who do not follow the study protocol, are not protected by randomization and thus may be affected by an imbalance in baseline variables (McNamee, 2009).

In practice, no matter how well designed and implemented a study, missing data are almost inevitable (Crutzen et al., 2013). In an RCT, missing data are more prevalent in outcome variables than in covariates, since data on covariates are usually collected at the time of enrolment. Different degrees of data incompleteness in these trials can occur as measurements may be available only at baseline or may be missed for one or several follow-up time-points. Incomplete outcome data in trials can lead to potential problems such as loss of efficiency due to reduced sample size and bias in the estimate of treatment effect due to differences between the observed and unobserved data (Horton & Lipsitz, 2001). For example, if missing data on an outcome are ignored and an analysis is

based only on the observed data, and these data are disproportionately from patients who are doing well in a new treatment arm and from patients who are doing poorly in the control arm, then the estimate of treatment effect could be overestimated. In addition, the benefits of randomization may be compromised; any statistical inferences, therefore, rely on additional assumptions. Further, a full dataset requires either imputation of missing values or modelling of unobserved data (European Medicines Agency, 2010).

Any analysis of RCTs with incomplete data is based on specific assumptions on the missing data mechanism, as detailed in chapter 2, such as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) (Rubin, 1976; Little & Rubin, 2002). For a valid analysis of RCTs with incomplete data, a transparent and plausible assumption on the mechanism of missing data is important (Little & Rubin, 2002). Since impacts of missing data on inferences of an RCT are difficult to assess and are closely associated with the question of what would have happened if all participants had completed that trial, it is important to try to gather information on the missing data, including the reason for its being missing, to make a justifiable assumption about the missing data.

As trials with missing data may not retain the balance of randomization, the basis for statistical inference is lost (Wright & Sim, 2003; Lewis & Machin, 1993) and there is no longer a statistical rationale to guarantee lack of bias for the estimation of the parameter and its associated confidence interval – even if the study is assumed to be free of other risks of bias, such as non-masked evaluation. Identification of the underlying missing data mechanism is important in order to carry out appropriate formal analyses of data with missing values; however, it is impossible to identify this mechanism with certainty based on the observed data alone (Fielding et al., 2009). Missing data should therefore be considered at the design, conduct and analysis stages of a trial (Molenberghs & Kenward, 2007; National Research Council, 2010; White et al., 2011). First, trialists should attempt

to minimize missing data in the first instance by following up all randomized subjects, even if they withdraw from an allocated intervention. Second, analysts should perform a primary analysis with a plausible assumption on the mechanism of missing data. Third, sensitivity analyses should explore the robustness of the results to a range of alternative plausible assumptions regarding missingness.

A few studies (Hollis & Campbell, 1999; Kruse et al., 2002; Wood et al., 2004; Gravel et al., 2007; Fielding et al., 2008) have examined practices regarding the use of the ITT principle and/or reporting and handling of missing data in RCTs published in general medical journals. Additionally, two studies have assessed the quality and application in RCTs in MSCs (Baron et al., 2005; Henschke et al., 2012). Baron et al. (2005) examined the proper use of the ITT approach and rate of missing data in 81 reports of superiority RCTs assessing structural outcomes in rheumatic diseases published between 1994 and 2003 in 10 general and 21 speciality journals. Henschke et al. (2012) evaluated trend over time (1980–2008) in quality of RCTs (n=157) of interventions for chronic low-back pain. The study was not limited to an evaluation of the use of ITT and the rate of missing data; in addition, several other study design characteristics associated with risk of bias were also considered. All these studies found many instances in which analyses were poorly defined and noted variation in practice regarding the ITT principle and the handling of missing data.

Hollis and Campbell (1999) examined 249 reports of RCTs published in 1997 in four medical journals (*BMJ, JAMA, Lancet* and *New England Journal of Medicine*) and reported that 48% of the reports claimed to perform an analysis according to 'ITT strategy', but conceded that the ITT analysis was not clearly described and applied in many instances. Kruse et al. (2002) evaluated a sample of 100 RCTs, published in 1999, that used the word 'intention-to-treat' or 'intent-to-treat' in their abstract. They reported that only 42% of the trials included all randomized subjects in their primary endpoint analysis.

In another study, Gravel et al. (2007) evaluated 403 reports of RCTs published in 2002 in ten medical journals and reported that 62% of the trials analysed their primary outcome on an ITT basis. However the reviewers found that only 39% of the identified trials actually analysed all subjects as randomized. The study also reported that 60% of the trials had at least some missing data and most of these trials (59%) excluded subjects with missing data from the primary analysis.

Wood et al. (2004) examined 71 trial reports published between July and December 2001 in *BMJ*, *JAMA*, *Lancet* and *New England Journal of Medicine* and reported that almost 89% had missing outcome data. Importantly, the review reported that outcomes were missing for more than 10% of participants in about half of the reviewed trials. The review further reported that 18% of the trials had more than 20% missing outcomes, and complete-case analysis was the most common approach in the primary analysis. Similar findings were observed in another review of the use of imputation methods to deal with missing quality of life outcomes in 61 RCTs published during 2005 and 2006 in the same medical journals (Fielding et al., 2008).

3.2.1 RCTs in MSCs

The number of RCTs on interventions for MSCs has increased steadily over the past decades (Koes et al., 2005; Henschke et al., 2012). Many RCTs in MSCs differ from the archetypal clinical trials that use placebo-controlled, double-blind methods (Akai, 2010). It is, for example, often difficult to utilize double-blind methods in pragmatic trials of clinical interventions for these conditions. For example, in a trial with comparison of standard treatment and standard treatment plus cognitive-behavioural training among patients with rheumatoid arthritis, imposing a double-blind method is not practical; only an open-label trial, where patients know about their treatment, is usually possible in such circumstances. In this situation, perceptions of disadvantages of one treatment over another may influence the patient's decision to continue in the trial. Additionally, the aims of treatment

for MSCs are mainly to reduce the burden of diseases and disability, and to improve health-related quality of life. Accordingly, the outcome measures are mostly related to participants' well-being or functional ability of the patients, not the quantification of biochemical or other laboratory data. Finally, owing to their chronic nature, many MSCs necessitate long-term trials, which are prone to loss to follow-up. Each of these features may predispose to missing values.

3.2.2 ITT analysis and missing outcome data in trials in MSCs

Baron et al. (2005) reviewed superiority trials published between 1994 and 2003 that assessed outcomes in rheumatic diseases (n = 81) and reported that a full ITT analysis – an analysis includes all randomized subjects in the analysis as randomized – was not applied in most (93%) of the identified trials. The researchers additionally examined the rate of missing data and found that about two-thirds of 63 reports in which missing data information had been reported had >10% participants with missing outcome data and approximately one-third had >20%. However, only a quarter of the reports reported statistical methods for handling the missing data.

Henschke et al. (2012) reported a study that examined the trend over time (1980–2008) in quality – based on study design characteristics associated with the risk of bias – of RCTs (n = 157) of interventions for chronic low-back pain. This evaluation was based on 11 criteria described by Koes et al. (2005), of which two (9 and 11) are closely linked to the ITT principle. Criterion 9 is fulfilled if the percentage of withdrawals and dropouts does not exceed 20% for short-term follow-up and 30% for long-term follow-up, and such dropouts must be described with reasons; more than one-third of the RCTs failed to fulfil the criterion most years during the study period. Criterion 11 is fulfilled if all randomized patients are analysed in the group they were allocated to by randomization for the primary effect measurement, minus missing values, irrespective of non-compliance and co-interventions; fewer than 60% of the RCTs fulfilled this criterion. Clearly the percentages

will fall further when considering both criteria 9 and 11 – which aligns with the strict definition of ITT whereby analysis extends to all individuals randomized whether data is lost or not.

3.2.3 Handling of missing outcome data in trials in MSCs

Baron et al. (2005) further investigated how missing data were handled in the reviewed trials, and reported that, after complete-case analysis, last observation carried forward (LOCF) was the most common method used to handle the missing data. The researchers also found that analysis methods that better address bias through missing data were limited to a very small fraction of the trials; one trial used mixed-effect statistical modelling. Since then, there have been several guidelines issued on prevention and treatment of missing data in RCTs (Food and Drug Administration, 2008; European Medicines Agency, 2010; National Research Council, 2010). No studies have been reported recently to assess the improvement in quality of reporting and handling of missing data in RCTs on MSCs.

3.3 Objectives

As mentioned in the background section, trials in MSCs are generally pragmatic in nature and hence issues of key relevance to the ITT principle, specifically adherence and compliance to treatment regimen and dropout through withdrawal and loss to follow-up, are prominent in this field, and possibly so to a greater extent than in other areas of clinical research. This study sought to examine the reporting of clinical trials in the musculoskeletal field in terms of the ITT principle and analysis issues relating to missing data.

Specifically, this study had the following objectives:

i. To describe the extent of reported dropout;

ii. To evaluate the extent of deviation from an ITT principle, and resulting loss to analysis in respect of the primary analysis;

iii. To evaluate the analytical methods used for handling the missing follow-up data in the main analysis;

iv. To evaluate the use of the sensitivity analyses performed to assess the robustness of inferences from the primary analysis to a range of alternative plausible missing data assumptions.

3.4 Methods

3.4.1 Selection of studies

3.4.1.1 Journal selection

Five journals (*Annals of the Rheumatic Diseases*, *Arthritis & Rheumatism*, *Journal of Rheumatology*, *Pain* and *Rheumatology*) were non-randomly selected as sources of RCTs in the areas of MSCs. The impact factors of journals (Thomson Reuters, 2011) were taken into consideration when the selection was performed. The impact factor of the selected journals ranged from 3.551 (*Journal of Rheumatology*) to 9.082 (*Annals of the Rheumatic Diseases*). The impact factor was taken in to account for two main reasons. Firstly, a high impact factor can be an indicator of higher methodological quality – a manuscript of a trial with weak methodological quality is less likely to be published in a journal that has a high impact factor than a manuscript of a trial with strong methodological quality (Lee et al., 2002). This argument is supported by a review of 469 RCTs published in 2007 in core clinical journals where trials published in higher impact journals had higher methodological quality compared with those published in lower impact journals in their reported design, conduct, and analysis (Bala et al., 2013). Secondly, acceptability of a published article – an article published in a journal that has a high impact factor is more likely to influence clinical practice and is more likely to be cited in future publications (Akl et al., 2009).

3.4.1.2 Eligibility criteria

The eligibility criteria were based on those used by Egbewale (2012). Only parallel-arm individually RCTs were included in this study. This restriction was imposed mainly because other trials like crossover trials and cluster RCTs follow a different design, analysis, and reporting strategy; moreover, the objectives of the thesis were centred on parallel-arm individually randomized clinical trials. More specifically, establishing the true ITT principle in the analysis of data is more troublesome in cluster RCTs than in individually RCTs because the strategy must be applied at the level of both the cluster and the individual, and because of challenging issues surrounding the recruitment process in cluster RCTs. Crossover trials depend on the absence of dropout and randomization is only in relation to the order of treatment assignment. So the issues are different in relation to application of the ITT analysis in these trials.

Primary reports of all such phase III clinical trials published in the aforementioned five journals over a two-year period were considered, but restricted to those reporting the primary outcome measure of the trial.

Additional exclusion criteria were:

i. Pilot/feasibility studies, as these mainly aim to demonstrate the feasibility and/or affordability of subsequently conducting a large similar study, rather than to detect a true between-group difference with sufficient power.

ii. Trials with a number of subjects randomized less than 50, as the small sample size could impose limitations on the possible methods of analysis.

iii. Publications based on an interim analysis (i.e. where the primary analysis was centred on an outcome measured at a time-point earlier than the designated primary endpoint).

iv. Extended follow-up studies (those that only considered outcomes beyond the primary endpoint).

v. Studies with survival outcomes, as standard survival models take into account lost to follow-up through non-informative censoring.

vi. Publications based on multiple trials.

3.4.1.3 Search strategy

A search was performed for all reports of RCTs published between 1st January 2010 and 31st December 2011 in these five journals. The advanced search option was used in each journal website with keywords 'clinical trial', 'randomization', 'randomisation', 'randomized', 'randomised', 'randomly', or 'random' in the titles or abstracts to identify relevant citations. This search strategy produced 405 relevant citations: 120 from *Annals of the Rheumatic Diseases*, 64 from *Arthritis & Rheumatism*, 70 from *Journal of Rheumatology*, 86 from *Pain*, and 65 from *Rheumatology*.

3.4.1.4 Selection and review process

Figure 3.1 illustrates the selection procedure and indicates the reasons for exclusions. An initial screening of titles, abstracts and keywords of all retrieved reports was performed to identify potentially relevant publications based on the eligibility criteria. The first screening resulted in exclusion of 206 reports. A copy of the full report of each selected publication was then accessed and screened. A further 108 reports were excluded during this second screening process. In total, 314 reports were excluded based on the eligibility criteria. Any additional tables and results relating to the trials referred to in the 91 reports finally selected for the review were obtained from the respective journal website. A list of the selected reports is provided in appendix 1.

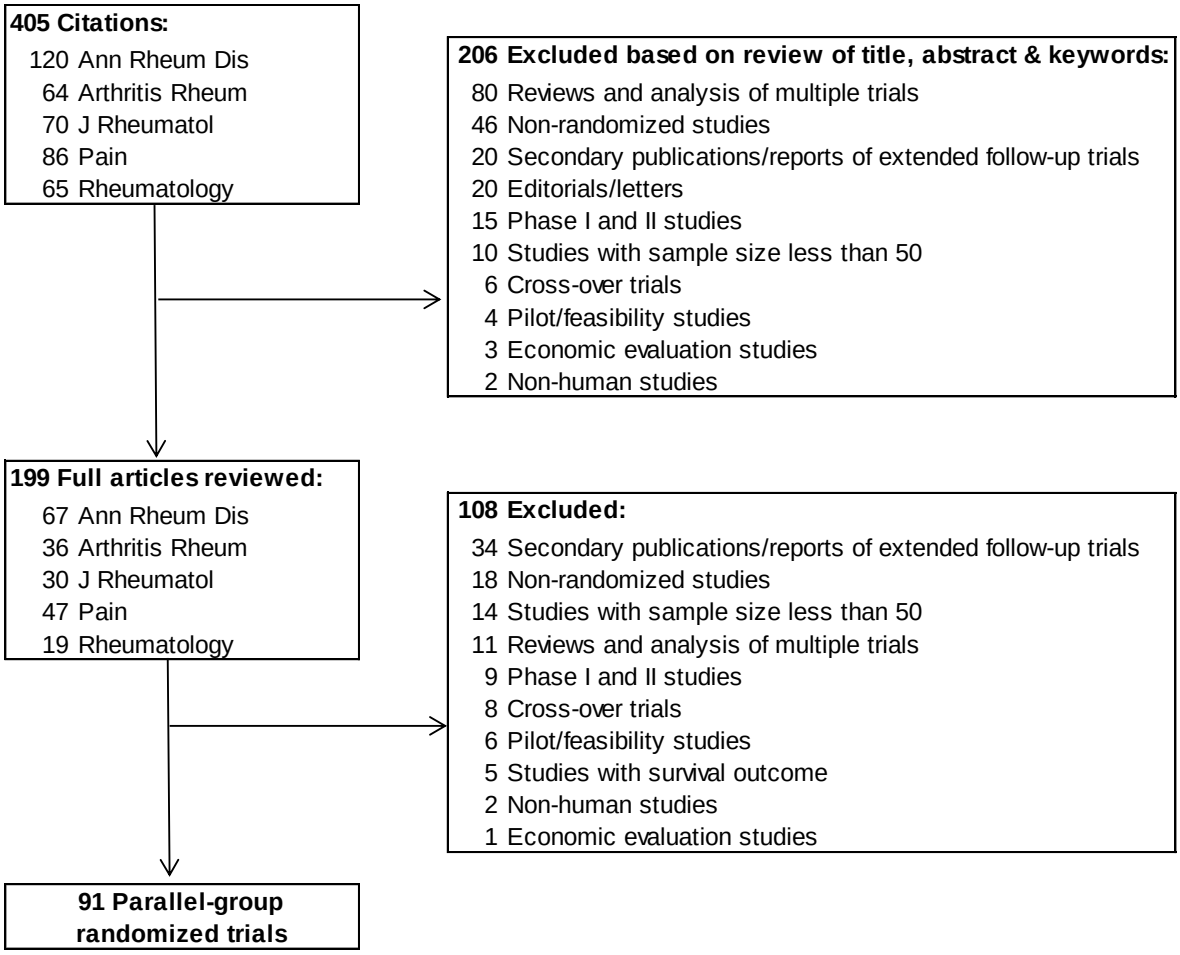


Figure 3.1: Identification of randomized trials from January 2010 to December 2011

3.4.2 Data extraction and management

A draft data extraction form was developed based on the CONSORT 2010 Statement (Schulz et al., 2010), the guideline on the prevention and treatment of missing data in clinical trials (National Research Council, 2010), and a research paper on the flow of participants in a randomized trial (Hopewell et al., 2011). The form was discussed with the supervisory team, and modified further until there was agreement on items and their operational definitions.

Data extraction mainly centred on the primary outcome (variable) at the primary endpoint (time). The primary outcome was identified from the definition given in the report (e.g. in the study objectives) or from the details on the sample size calculation. If more than one

primary outcome was reported, the first one listed was used. If, in the case of multiple follow-ups, the primary endpoint was not explicitly identified, it was taken to be the final measurement. The decision on the choice of the primary endpoint has been taken because MSCs are commonly long-standing, and the aim is usually to address long-term pain and functional difficulties. Further, a primary analysis was defined as the analysis of the primary outcome at the primary endpoint.

Data were extracted, using the proforma (Appendix 2), from each of the 91 eligible trial reports. Information was obtained on basic characteristics of the trials, participants' exclusions and withdrawals, methods used to handle the missing outcome data, statistical analyses performed, and sensitivity analyses performed.

Specifically, reports were scrutinized for: number of centres involved, number of arms involved, details on primary outcome variable (including the type of variable used for measurement and analysis; sometimes a continuous primary outcome is categorized for the purpose of a primary analysis), number of follow-up visits after baseline visit, and details on sample size (including calculated sample size and whether adjustment for attrition in the calculation was done). When the information on sample size calculation was not given in an article, the details were obtained from a previously published paper on the same trial if available. In addition, data were extracted on number of subjects randomized, allocated to each group, receiving the allocated intervention, not receiving the allocated intervention, completing the primary outcome measure at the primary endpoint, and included in the primary analysis. Apart from the numbers, information was also collected on the reasons for not completing the primary outcome assessment at the primary endpoint and for exclusion of participants from the primary analysis. In addition, the method used, if any, to deal with missing data was identified. Information was also gathered on handling of treatment crossover, or other protocol violations (e.g. subjects

may not have followed the treatment procedure properly or may have taken other medication along with their assigned treatment).

3.4.2.1 Calculation of dropout rate

Dropouts were subjects who did not complete the primary outcome assessment at the primary endpoint, whereas completers were those who completed the assessment. Dropouts include individuals lost to follow-up (through non-response) and those not followed up due to study protocol violations such as ineligibility or treatment crossover. Dropout rate was calculated as the difference between the number randomized and the number remaining in the trial (completers) at the primary endpoint, divided by the number randomized. In trials with repeated follow-ups, the dropouts were classified as either “early dropouts”, defined as subjects who did not complete *any* follow-up assessment, or “late dropouts”, defined as those who completed *at least one* follow-up assessment prior to dropping out. The early and late dropout rates were calculated based on the number of subjects randomized.

3.4.2.2 Classification of analysis strategies

The analysis strategy used in the reviewed reports was categorized as: full ITT (FITT) analysis, partial ITT (PITT) analysis, complete-case analysis (CCA), PP analysis, or AT analysis. The definition of each category is given in table 3.1. FITT is an analysis of data as randomized, and includes data on all randomized subjects through either imputation or modelling of any missing data. PITT denotes an analysis restricted to a subset of the full ITT sample where the sub-sample excludes early dropouts (in trials with repeated follow-ups). The purpose of this classification is to highlight the exclusion of early dropouts from the primary analysis. In trials with a single follow-up, the exclusions lead to a CCA as there is no scope for further follow-up data. White et al. (2012) argue that including all

randomized subjects in an analysis of an outcome with missing data is insufficient; suggesting one should also consider an appropriate method to handle the missing data.

Table 3.1: Classification of analysis strategy used in trial reports

Statistical analysis strategy	Explanation
Full ITT (FITT) analysis	All randomized subjects included in the analysis and analysed as randomized.
Partial ITT (PITT) analysis	Analysis includes all randomized subjects except those who did not provide *any* follow-up data on the primary outcome in trials with repeated follow-ups.
Complete-case analysis (CCA)	Analysis includes only those randomized subjects who completed the primary outcome measurement at the primary endpoint. i.e., this analysis excludes subjects with missing data at the primary endpoint.
As-treated (AT) analysis	All subjects analysed as treated, regardless of the treatment to which they were assigned.
Per-protocol (PP) analysis	Analysis includes subjects who completed the trial in full accordance with the study protocol.

3.4.2.3 Definition of loss to analysis

The exclusion of randomized subjects from the primary analysis was referred to as 'loss to analysis' (Deo et al., 2011) and was calculated as the difference in the number

randomized and the number included in the primary analysis. Deviation from the FITT analysis generally results in loss to analysis. The reason for loss to analysis can be either loss of participants in the analysis with lack of measured or imputed outcome, or exclusion of participants with a measured outcome because of certain reasons (e.g. non-adherence to treatment protocol), or both. For example, in PP analysis, loss to analysis includes completers who were excluded for reasons of protocol violation in addition to dropouts, whereas in a PITT analysis, it includes only early dropouts.

3.4.2.4 Classification of methods to handle missing data

The method used, if any, to deal with missing data was classified as:

i. Methods that lead to listwise deletion of subjects with missing data, and moment-based methods, such as generalized estimating equations (GEE).

ii. Single imputation methods, such as baseline observation carried forward (BOCF), last observation carried forward (LOCF), worst observation carried forward (WOCF), non-responder (i.e. treatment failure) imputation (NRI), regression method or linear extrapolation method. In NRI, dropouts are assumed to be non-responders regardless of their prior response status at the time of dropout, where an outcome is analysed on a dichotomous or categorical scale.

iii. Multiple imputation (MI) – a method to overcome certain limitations (e.g. risk of underestimating the variance of treatment effect) of single imputations.

iv. Statistical models that can include all randomized subjects without imputation of missing values. For example, full-information maximum likelihood (FIML) based methods, such as mixed-effects model for repeated measures (MMRM).

3.4.2.5 Data accuracy

To ensure accuracy, the process of data extraction was repeated after several months, blind to the outcome of the initial data extraction. Any discrepancies were reconciled through a third review of corresponding reports. Another reviewer (RO) then independently verified the corrected data against a random selection of 20 reports (22%), and identified discrepancies in less than 1% of total data points. These discrepancies in the data extraction were resolved by discussion.

3.5 Results

3.5.1 Characteristics of included trials

A description of the 91 trials included in the review is presented in table 3.2. More than half (57%) were multicentre trials. Three-quarters were two-arm studies, while 10% had more than three study arms. Even though 72 trials had a numerical primary outcome measure, 13 of them categorized this continuous outcome measure; for example, one of the included trials (Russell et al., 2011) defined favourable/clinically-improved outcome as 30% reduction in pain visual analogue scale score from baseline to the end of the treatment period. In 80 (88%) trials, there was assessment of the primary outcome at two or more follow-ups; hence, at least one intermediate outcome measurement was available before the assessment at the primary endpoint.

Table 3.2: Description of the selected trials (n=91)

Description of the trials	No. of trials (%)
Journal	
Annals of Rheumatic Diseases (IF: 9.082)	31 (34.1)
Arthritis & Rheumatism (IF: 8.435)	16 (17.6)
Journal of Rheumatology (IF: 3.551)	11 (12.1)
Pain (IF: 5.355)	27 (29.6)
Rheumatology (IF: 4.171)	6 (6.6)
Year of publication	
2010	38 (41.8)
2011	53 (58.2)
Multicentre trials	
Yes	52 (57.1)
Number of subjects per trial[1]	
<100	36 (39.6)
100–499	41 (45.0)
500 and above	14 (15.4)
Number of arms per trial	
2	67 (73.6)
3	15 (16.5)
>3	9 (9.9)
Type of primary outcome measure	
Categorical	19 (20.9)
Numerical[2]	72 (79.1)
Number of follow-up assessments	
Single	11 (12.1)
Repeated	80 (87.9)

IF – Impact factor; [1]Number of subjects randomized; [2]13 trials analysed these outcomes as categorical

Nineteen (21%) trials did not present any formal sample size calculation (Table 3.3). Among the 72 trials that detailed the sample size calculation, only 28 (39%) made adjustment for attrition. In 18/72 (25%) trials, the number of subjects randomized was less than the calculated sample size (the shortfall ranged from 1% to 53%; median 5%). The number of subjects randomized in the identified trials ranged from 51 to 1025, with a median of 140. The number of subjects remaining in the trials at the primary endpoint ranged from 32 to 786, with a median of 107 subjects. The number of subjects included in the primary analysis ranged from 32 to 1025, with a median of 115.

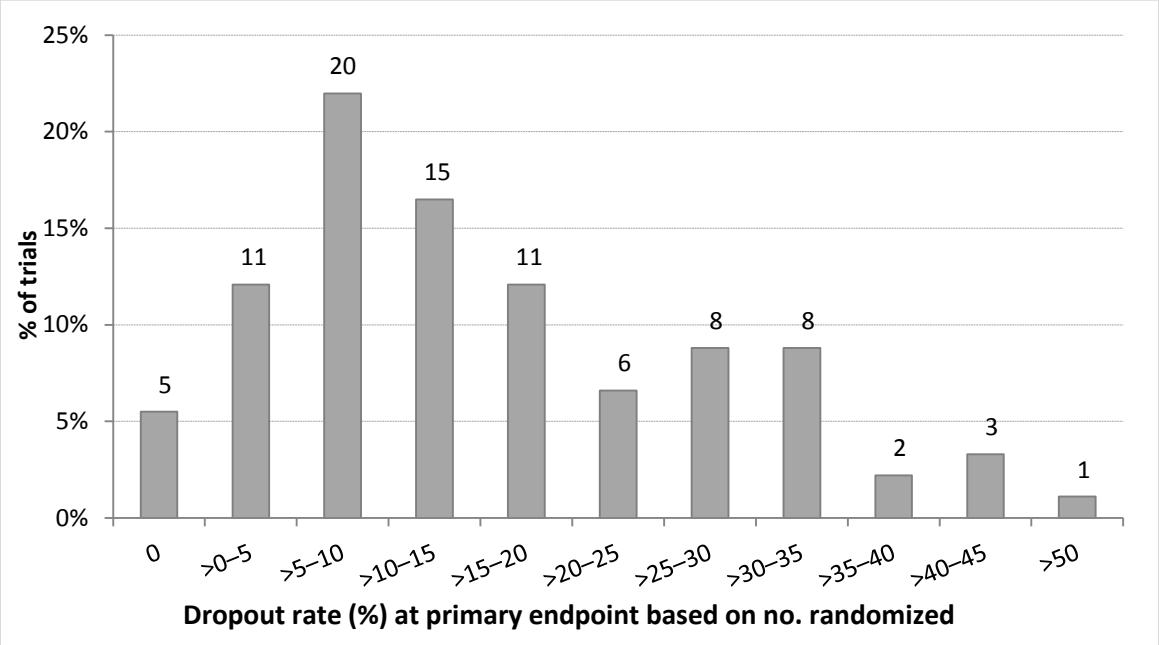
Table 3.3: Size of the trial

Details	No. of trials reporting the details	Number of subjects	
		Median (IQR)	Min – Max
Calculated sample size	72[a]	136 (74–366)	38–1025
Number of subjects randomized	91	140 (76–306)	51–1025
Number of subjects completed the trial	90[b]	107 (66–247)	32–786
Number of subjects included in the primary analysis	91	115 (71–261)	32–1025

[a]In 28 trials this included an adjustment for attrition; [b]It was clear from the remaining one article that it had few missing data

3.5.2 Dropouts

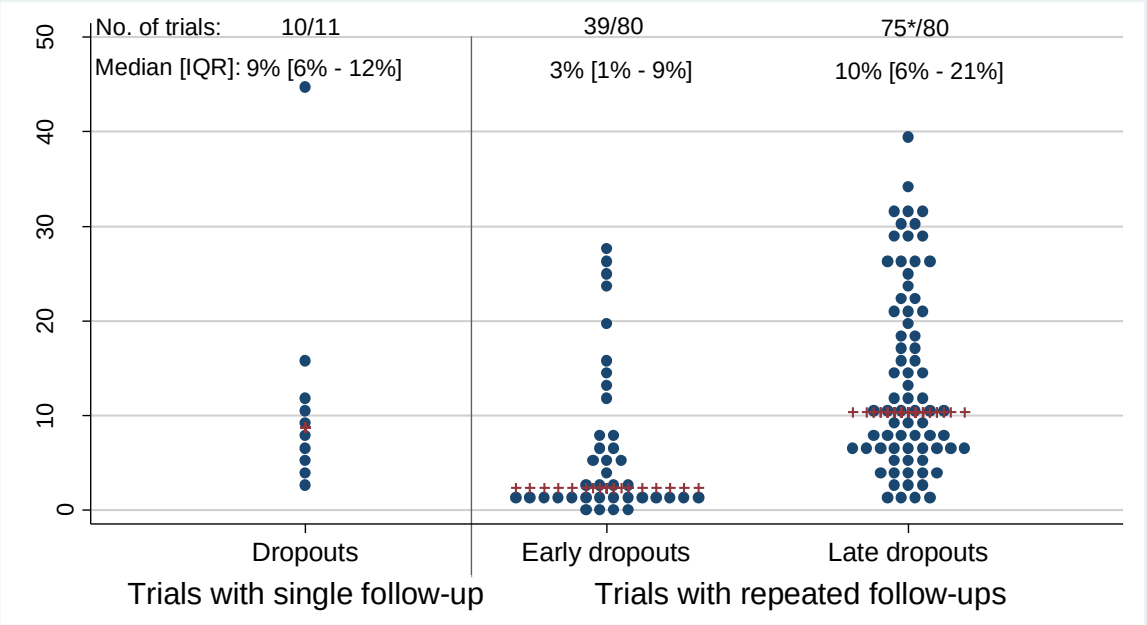
All but one trial reported the number or proportion of subjects with missing data at the primary endpoint. In the one exception it was clear from the report that there were missing values at the primary endpoint. However, arm-specific details were not clearly reported in 12 trials. Figure 3.2 displays the percentage of trials with various levels of dropouts. Eighty-six trials (95%) had some subjects with missing outcome data at the primary endpoint. The median dropout rate was 12% (IQR 7% to 24%; range 0% to 51%). Fifty-four (60%) trials had more than 10% dropouts and 29 (32%) had more than 20% dropouts based on number of subjects randomized.



The number above each bar indicates the number of trials; *One trial did not report the dropout rate

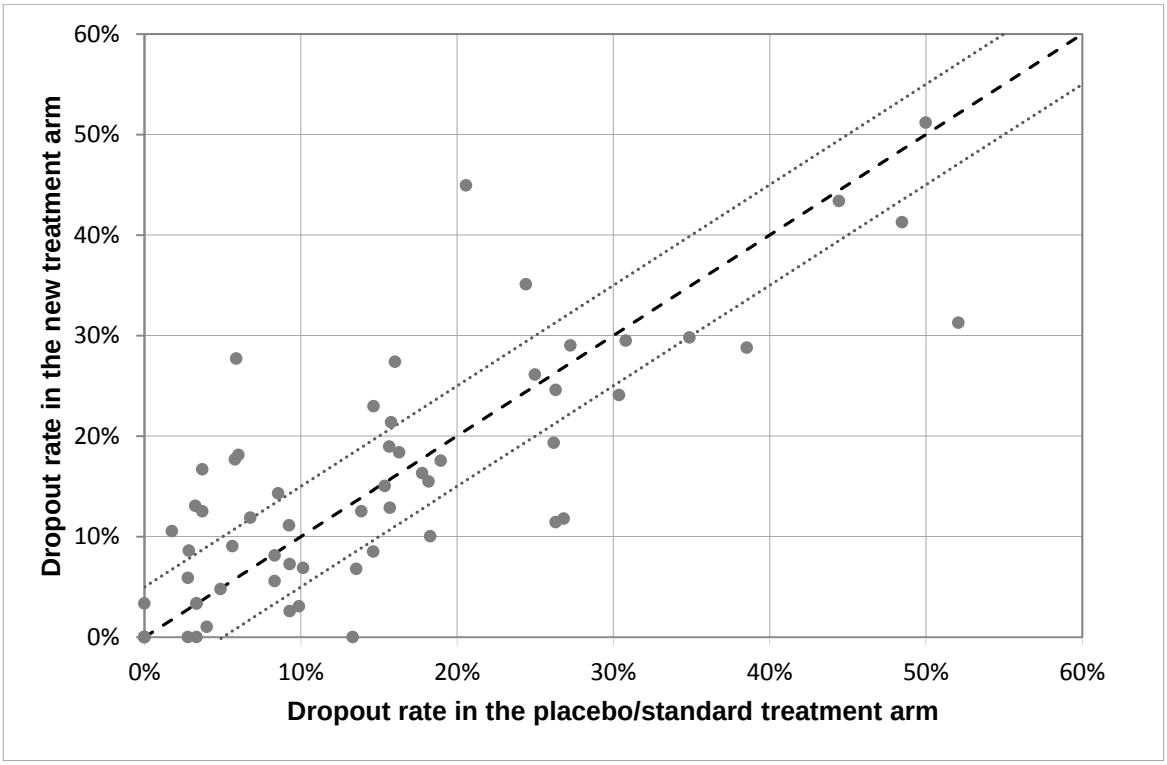
Figure 3.2: The distribution of the 90 trials based on the percentage of dropouts

The distribution of the dropout rates by number of follow-ups is shown in figure 3.3. Among 11 trials with a single follow-up, 10 (91%) trials reported dropouts (median [IQR] dropout rate of 9% [6%, 12%]). Among 80 trials with repeated follow-ups, 39 (49%) trials reported early dropouts (median [IQR] dropout rate of 3% [1%, 9%]), and 75 (94%) trials reported late dropouts (median [IQR] dropout rate of 10% [6%, 21%]).



*One trial did not report the dropout rate; +indicates median dropout rate

Figure 3.3: Dropout rate at primary endpoint by number of follow-ups



The bold dashed line indicates line of perfect equality, and the upper and lower dashed lines indicate +/- 5% from perfect equality

Figure 3.4: Dropout rate at the primary endpoint between arms (n=61)

Figure 3.4 displays the pattern of dropout rates between arms as a scatter plot. Among 67 trials with two arms, four trials had no dropout, 28 (42%) trials had nearly equal dropout rates between the arms (the difference in dropout rates between arms was less than 5%), 15 (22%) had higher dropout rate in the new treatment arm, 14 (21%) had higher dropout rate in the placebo/standard treatment arm, and the remaining six trials did not report the arm-specific dropout details.

3.5.3 Analysis strategy and loss to analysis

Table 3.4 indicates the analysis strategy followed in the primary analysis of the trials. FITT analysis was performed in 34 (37%) trials; in four trials (one with a single follow-up and three with repeated follow-ups), there were no missing outcome data at the primary endpoint and in the remaining 30 trials (one with a single follow-up and 29 with repeated follow-ups), all randomized subjects were included in the analysis through either imputation of missing values or an analysis that accommodates missing data.

In 24 (26%) trials (seven with a single follow-up and 17 with repeated follow-ups), only completers were included in the analysis. In detail, all dropouts were excluded from nearly two-thirds of trials with a single follow-up (median [IQR] loss to analysis of 10% [8%, 15%]) and one-fifth of trials with repeated follow-ups (median [IQR] loss to analysis of 8% [6%, 17%]).

Among 10 trials that reported deviations from allocated treatment (i.e. crossover of treatment) after randomization, one trial (Rubbert-Roth et al., 2010) performed the primary analysis on an AT basis. Protocol violations were reported in 22 trials, and four of them (two each with a single follow-up and repeated follow-ups) followed a PP strategy as the primary analysis.

Table 3.4: Analysis strategy followed in the primary analysis. Data are counts (%)

Analysis strategy	Trials with single follow-up	Trials with repeated follow-ups	Total
Intention-to-treat (ITT)	2 (18.2)	60 (75.0)	62 (68.1)
Full ITT	2 (18.2)	32 (40.0)	34 (37.4)
Partial ITT	n/a	28 (35.0)	28 (30.7)
Complete-case analysis	7 (63.6)	17 (21.3)	24 (26.4)
As-treated analysis	0 (0.0)	1 (1.2)	1 (1.1)
Per-protocol analysis	2 (18.2)	2 (2.5)	4 (4.4)
Total	11 (100.0)	80 (100.0)	91 (100.0)

n/a: not applicable

In another 28 (35%) of the 80 trials with repeated follow-ups, a PITT analysis was performed (excluding early dropouts). Among these trials, median [IQR] dropout rate was 21% [11%, 30%] and median loss to analysis was 2% [1%, 6%]. The terms 'ITT' and 'modified ITT' were used interchangeably across these trials to denote the analysis strategy. As noted in table 3.5, nine trials failed to give a clear description of the analysis strategy that was followed in the trial report. Various descriptions were used to define the analysis strategy among the remaining 19 trials (Table 3.5). The descriptions varied in relation to several reasons for the exclusions from the analysis; patients were ineligible because they were mistakenly randomized, patients did not start the intervention or did not complete the entire course, patients did not provide a baseline assessment, or patients did not complete any post-baseline assessment.

Table 3.5: Description on analysis strategy provided in the 28 trial reports with classification 'partial ITT'

Analysis sample	Number of reports (n=28)	Loss to analysis (%)	Reports
Not defined	9	1%–20%	Atchia et al., 2011; Beaudreuil et al., 2011; Brien et al., 2011; Christiansen et al., 2010; Keefe et al., 2011; Machold et al., 2010; Oldenmenger et al., 2011; Schmidt et al., 2011; Zangi et al., 2011
All randomly assigned patients who received at least one dose of study medication	10	0.5%–7%	Bliddal et al., 2011; Emery et al., 2010; Furie et al., 2011; Griffiths et al., 2010; Jones et al., 2010; Katz et al., 2011; Molsberger et al., 2010; Pauer et al., 2011; Strand et al., 2011; Taylor et al., 2011
All randomly assigned patients who received at least one dose of study medication, and had a baseline as well as at least one post-baseline values for the primary outcome	5	0.5%–14%	Baranauskaite et al., 2011; Branco et al., 2010; Tak et al., 2011; Kim et al., 2011; Seibold et al., 2010
All randomly assigned patients with at least one efficacy assessment after randomization	2	2% & 16%	Kravitz et al., 2011; Navarro-Sarabia et al., 2011
All randomly assigned patients who met the inclusion criteria and were followed up	1	12%	Alavi et al., 2011
All randomly assigned patients who met the study eligibility criteria, received at least one dose of study medication, and had a baseline as well as at least one post-baseline values for the primary endpoint	1	3%	Tanaka et al., 2011

3.5.4 Handling of dropouts: imputation strategy

Among the ten trials with a single follow-up that reported dropouts, nine excluded the dropouts from analysis, while the remaining one trial employed BOCF in the analysis. Among the 39 trials with repeated follow-ups that reported early dropouts, 36 (92%) excluded those dropouts from the analysis (28 followed PITT, 7 followed CC, and one followed AT analysis), and the remaining three each employed BOCF, MI or MMRM to handle the missing outcome data.

Table 3.6 shows the methods used to handle the missing values occurring after at least one follow-up assessment among the trials with repeated follow-ups. Among the 75 trials with late dropouts, 26 (35%) trials did not use any kind of imputation, 44 (59%) trials performed some sort of single imputation, two (3%) trials performed MI to replace missing values, and the remaining three trials did not provide the relevant details. Among the trials that did not use any imputation approaches (n=26), eight trials used analysis methods that made full use of all available data (MMRM in six and GEE in two) and the remaining 18 trials excluded the dropouts from the primary analysis without considering the availability of the outcome data measured at an earlier follow-up time. It can be seen that LOCF was the most frequently used imputation approach, used in 23 (31%) trials. The BOCF approach was followed in another four trials. In 12 (16%) trials, subjects with missing values were dealt with through non-responder imputation; in two of these trials, the outcome was measured on a continuous scale but analysed as categorical.

Table 3.6: Methods used to handle late dropouts, who had completed at least one follow-up assessment. Data are counts (%)

Method used in primary analysis	Percentage of late dropouts			
	>0–10%	>10–20%	>20%	Total
No imputation				
Excluded	10 (29.5)	6 (33.3)	2 (9.1)	18 (24.0)
MMRM	2 (5.9)	1 (5.6)	3 (13.6)	6 (8.0)
GEE	1 (2.9)		1 (4.6)	2 (2.7)
Single Imputation:				
LOCF[1]	7 (20.6)	7 (38.8)	8 (36.3)	23 (30.7)
NRI	7 (20.6)	2 (11.1)	3 (13.6)	12 (16.0)
BOCF	1 (2.9)		3 (13.6)	4 (5.3)
Regression imputation	1 (2.9)	1 (5.6)		2 (2.7)
LOCF + NRI[2]		1 (5.6)		1 (1.3)
LOCF + WOCF[3]			1 (4.6)	1 (1.3)
Linear extrapolation	1 (2.9)			1 (1.3)
Multiple imputation	2 (5.9)			2 (2.7)
No details	2 (5.9)		1 (4.6)	3 (4.0)
Total	34 (100)	18 (100)	22 (100)	75 (100)

MMRM – mixed model for repeated measures; GEE – generalized estimating equations; LOCF – last observation carried forward; NRI – non-responder imputation; BOCF – baseline observation carried forward; WOCF – worst observation carried forward; [1]dropout rate was not reported in one trial; [2]trial in which subjects dropping out were treated as non-responder when dropout due to adverse events or lack of effectiveness, otherwise imputed with LOCF; [3]trial imputed with WOCF when dropout due to adverse events or inefficacy, otherwise imputed with LOCF

Figure 3.5 depicts the approaches to missing data followed in 40 (50%) longitudinal studies with more than 10% late dropouts. Importantly, subjects with missing outcomes were completely excluded from eight (20%) trials; in the remaining 32 trials, 15 used LOCF to accommodate subjects with missing outcome data. Further, MMRM was employed only in four (10%) trials, and MI was not performed in any of these trials.

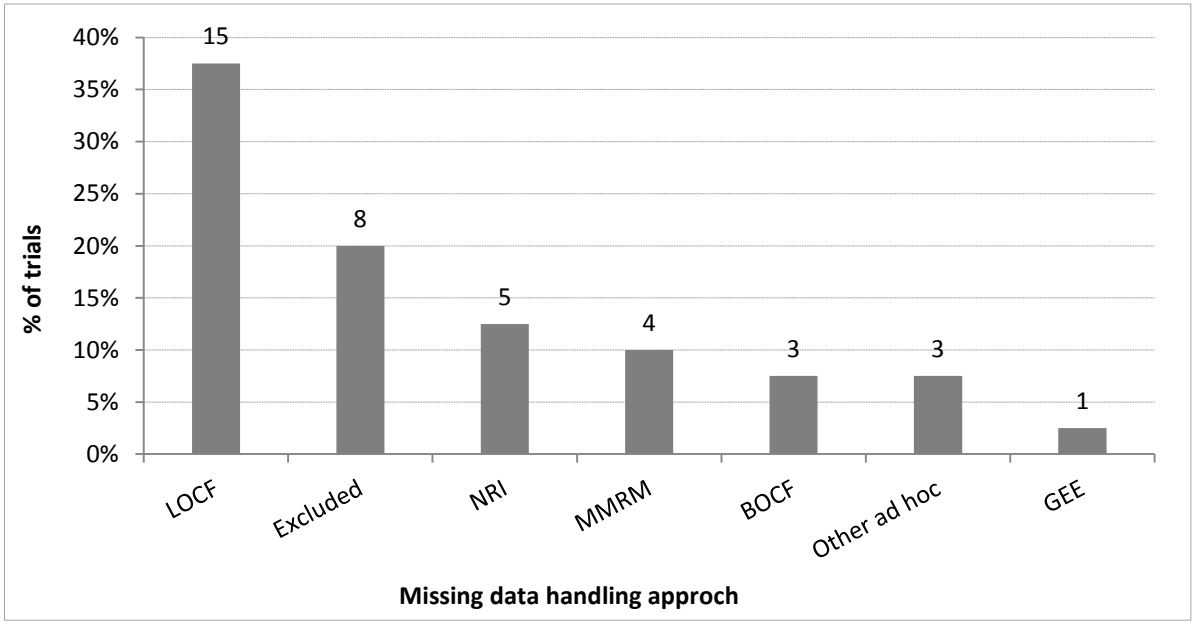


Figure 3.5: Handling of missing data in trials with >10% late dropouts (n=40; detail was not reported in one trial)

3.5.5 Sensitivity analysis and cautionary notes on missing data

Eighteen (21%) out of 86 trials with missing outcome values at the primary endpoint reported a sensitivity analysis to assess the robustness of inferences from the primary analysis to a range of alternative plausible assumptions regarding missing data (Table 3.7). The sensitivity analyses were performed in trials with relatively high proportions of missing data (median 24%; IQR 17%, 33%). Either exclusion of subjects with missing data (i.e. listwise deletion) or a single imputation method was the designated sensitivity analysis. Very few trials (6/18; 33%) presented the results of their sensitivity analysis, while the others just reported that a sensitivity analysis had been performed and indicated that the findings from the primary analysis were supported by those of the sensitivity analysis.

We further reviewed all 54 trial reports with 10% or more missing data to determine how the authors addressed the uncertainty due to missing data in their interpretation of results. Among those trials that did not report a sensitivity analysis, none attempted to highlight

the uncertainty around their findings due to the missing data, apart from a few instances in which dropout was briefly identified as a limitation of the study.

Table 3.7: Description of trials that performed a sensitivity analysis for missing data

Study	No. of subjects	% of dropouts	Primary analysis	Sensitivity analysis(es)
Verstappen et al.	268	5.6	CCA	Treated as responder, NRI
Genevay et al.	61	6.6	LOCF with MMRM	BOCF
Beaudreuil et al.	70	11.4	MI	CCA
Gabay et al.	162	14.2	LOCF	Linear interpolation
Lumley et al.	88	17.0	CCA	Latent growth curve modelling
Ginzler et al.	370	17.3	LOCF	CCA
Kravitz et al.	307	18.9	MMRM	MI with MMRM
Branco et al.	884	21.6	LOCF	BOCF
Machold et al.	389	22.1	LOCF	Treated as a responder
Turner et al.	191	25.7	CCA	MI
Keefe et al.	116	27.6	MMRM	Pattern-mixture models, CCA
Arnold et al.	1025	30.1	BOCF	CCA, LOCF, MMRM
Furie et al.	826	30.3	Non-responder	LOCF, CCA
Griffiths et al.	90	32.2	LOCF	CCA
Thorn et al.	83	34.9	MI with MMRM	MMRM
Russell et al.	548	39.1	BOCF	LOCF
Bliddal et al.	96	41.7	BOCF with MMRM	CCA
Hewlett et al.	168	50.6	CCA	MI

3.6 Discussion

3.6.1 Overall summary

This review of recently published trial reports in five major musculoskeletal journals illustrates current practice relating to the ITT principle and the handling of missing outcome data in the primary analysis of RCTs for these clinical conditions. Missing outcome data were of concern in most of the reviewed trials; nevertheless 68% trials in this review analysed all randomized subjects who had at least one follow-up assessment. In particular, a FITT (with emphasis on endpoint analysis of all randomized subjects) was performed in 37% and a PITT (based on analysis of randomized subjects excluding individuals displaying early dropout in trials with repeated follow-ups) was performed in another 31%. However, since most trials failed to use appropriate statistical methods to account for missing data, it is likely that descriptive estimates and hence, the inference on treatment effect were biased, given that missing data are likely to differ from reported data.

This review noted inconsistency in reporting baseline differences of participants when loss to analysis is substantial; several trials evaluated the differences in baseline characteristics between arms based on number of subjects randomized rather than on number analysed despite a discrepancy between these numbers. This oversight fails to locate (and hence may fail to adjust for) any imbalance in the baseline characteristics between arms in the analysis dataset. Further, most trials failed to justify the assumption made during the analysis. Almost no trials considered the possibility of MNAR; moreover, many of the trials failed to adopt a method that is valid under the MAR assumption, which is a recommended neutral starting point in many settings (Molenberghs & Kenward, 2007; National Research Council, 2010; White et al., 2011). Importantly, sometimes the applied method (e.g. LOCF) may not even be valid under MCAR (European Medicines Agency, 2010; National Research Council, 2010). Some researchers – e.g. Navarro-Sarabia et al.

(2011) – used LOCF with an expectation that this approach can provide a conservative estimate of treatment effect; but this may not be true in some situations (European Medicines Agency, 2010). A similar finding was observed in sensitivity analyses as well.

The findings of this study are comparable with previous studies (Hollis & Campbell, 1999; Kruse et al., 2002; Wood et al., 2004; Gravel et al., 2007) that reviewed trials published in general medical journals. It is a concern that progress has not been made in reducing the large proportion of trials that are inappropriately analysed, and which therefore may be prone to erroneous estimates and conclusions.

3.6.2 Quality of reporting

When compared to trials in the previous studies (Hollis & Campbell, 1999; Wood et al., 2004; Baron et al., 2005; Gravel et al., 2007), the overall reporting quality of the trials in this review was higher. Most of the trials in this review followed the CONSORT statement (Moher et al., 2001; 2010) in reporting the trial results. The 2010 CONSORT flow diagram demands detailed information on arm-wise progress through various phases of an RCT. It explicitly requires reporting of arm-wise numbers of subjects: assessed for eligibility; declining to participate (with reasons); randomized; allocated to intervention; not receiving the allocated intervention (with reasons); lost to follow-up (with reasons); discontinued the intervention (with reasons); and excluded from the analysis (with reasons). In this review, five trial reports failed to provide the flow diagram. Another seven trials failed to report arm-wise detail on number of subjects randomized, but provided the number of eligible subjects who had started the intervention. This clearly limited the scope for calculating arm-wise dropout rate, and for assessing the association of assigned interventions with exclusions due to ineligibility and failure to start the allocated intervention.

3.6.3 Importance of collecting data on all randomized subjects

As Lavori et al. (2008) specified, it is important to take careful steps to minimize missing data in the trial design and data collection. It helps to reduce the need to use unverifiable assumptions about the missing data and thus minimize problems in inferential analyses, especially those that flow from misspecification of missing data assumptions in the analysis. Trialists should give greater attention to missing data that may be influenced by unobserved data. For example, study participants may wish to miss a hospital visit for outcome assessment when such a visit may be either difficult for those who are too ill or unnecessary for those who are feeling well. Researchers may prefer a home visit instead of requesting the participants to make a hospital visit. Another important point is to gather information on the reasons for missing data and on patient characteristics that can predict it. This is highlighted in the National Academy of Science (NAS) report (National Research Council, 2010), which contains recommendations regarding the prevention and treatment of missing data in clinical trials. Recommendations 3–5 specifically suggest obtaining more information on post-withdrawal data (Table 3.8). The additional information may help to justify an ignorable missing data and allows adjustment for bias in treatment effect through appropriate methods of analysis. Additionally, trialists should consider the availability of secondary sources to obtain outcome data information when there is data missing from the primary source. For example, information on outcome data was collected from general practice notes in one trial (Verstappen et al., 2010): data on disease-modifying antirheumatic drug use were obtained from GPs on 23 patients who discontinued the study and did not want to attend for further assessments. The accuracy of the data from a secondary source may be questioned; however, a trade-off between the accuracy and having missing data is required.

Table 3.8: Recommendations 3–5 of the NAS report on missing data

#	Recommendations
3	Trial sponsors should continue to collect information on key outcomes on participants who discontinue their protocol-specified intervention in the course of the study, except in those cases for which a compelling cost-benefit analysis argues otherwise, and this information should be recorded and used in the analysis.
4	The trial design team should consider whether participants who discontinue the protocol intervention should have access to and be encouraged to use specific alternative treatments. Such treatments should be specified in the study protocol.
5	Data collection and information about all relevant treatments and key covariates should be recorded for all initial study participants, whether or not participants received the intervention specified in the protocol.

3.6.4 Power calculation in anticipation of dropouts

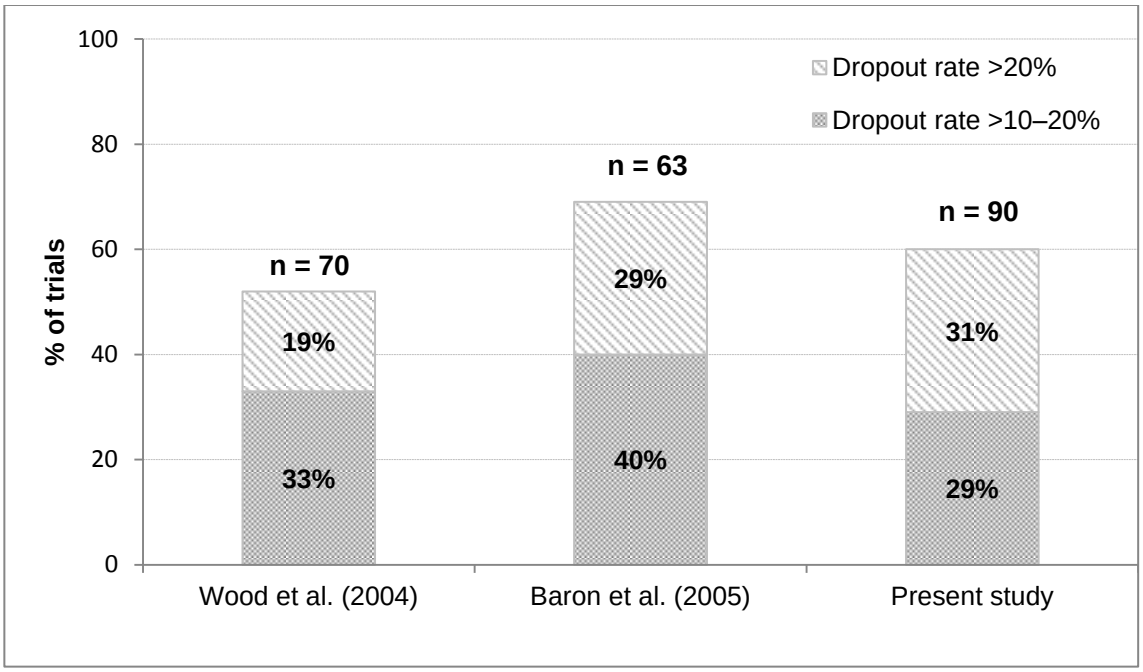
In the review, we found that one-fifth of trials failed to report a formal sample size calculation, contrary to the requirements of the CONSORT statement (Moher et al., 2010). This is important because specification of a primary outcome variable and primary endpoint guards against changing the planned outcome and placing undue emphasis on an outcome that was not the original primary outcome. Additionally, these calculations alert readers to potential problems like loss of power due to problems with participant recruitment and retention in RCTs. Most of the RCTs in this review did not meet their sample size targets. Specifically, 25% (18/72) of trials reporting a sample size calculation failed to achieve adequate numbers at randomization and 62/72 did not meet the target set for the primary endpoint.

As found in this review, trials often fail to recognize the importance of adjustment for attrition in sample size calculations in order to retain sufficient statistical power to detect a true treatment effect. Only a limited number of trials in the review (28/72) reported a kind of adjustment for anticipated dropout rate – inflating sample size in proportion to the dropout rate. In a quarter of the trials in the review (18/72), the number of subjects randomized was less than the calculated sample size. These findings raise many fundamental questions, such as: (i) Why did a large proportion of trials fail to perform and/or report the adjustment for attrition?; (ii) Do those methods not leading to listwise deletion protect against loss of power?; (iii) Does inflating sample size protect against the loss of statistical power due to attrition? A detailed discussion around these issues will be carried out in chapter 6.

3.6.5 Dropout rate

Dropout was not limited to a small proportion of trials. Almost 95% of trials in this review reported dropouts, which ranged from 1% to 51% of randomized subjects. The average dropout rate was a little over 10%. Compared to previous studies (Wood et al., 2004; Baron et al., 2005), which reviewed RCTs published nearly a decade ago, the present study did not observe any major change in the level of missing data. As seen in figure 3.6, Wood et al. (2004) reported that more than a half of reports had a dropout rate >10% and 19% had >20%. In Baron et al.'s (2005) study, 69% of 63 reports that clearly provided a missing data proportion had a dropout rate >10% and 29% had >20%. The amount of dropouts in this study may vary if one considers those reports (n = 18) that did not report the rate of missing data. The higher percentage is perhaps attributable to the inclusion criteria of that study, as it did not distinguish reports that were based on the primary outcome variable from those that were not. Usually, secondary outcome variables are less focused and more prone to missing outcomes in RCTs (Hopewell et al., 2013). Indeed, the present study also identified that a high proportion (60%) of trials had more than 10%

dropouts and 31% had >20%. The high proportion of trials with a significant number of dropouts is concerning.



n = number of trials reviewed

Figure 3.6: Comparison of reviews in relation to percentage of trials with various levels of dropout rates

The benefits of randomization in RCTs may be threatened if not all randomized participants are included in the analysis. Exclusion of participants without any follow-up outcome was the most common approach among the trials in this review. Such excluded participants, as expected, were more prevalent in trials with a single follow-up compared to trials with repeated follow-ups. Trialists should therefore consider collecting interim measures of the outcome, which may increase the number of subjects with at least some outcome data, and makes an MAR assumption more plausible (Wood et al., 2004). Still, nearly half of trials with repeated follow-ups in this review reported early dropout after randomization; nearly a quarter of them had >10%. Reasons for such dropouts varied across trials: intervention not received/completed, absence of baseline assessment, withdrawn consent, and loss to follow-up due to other reasons (such as location change)

were the major reason for early dropouts. A small proportion was due to practitioners' decisions such as exclusion of randomized subjects who were subsequently identified as not eligible. In most situations, it was rather unclear how the assigned intervention influenced the dropouts.

Late dropouts in trials limit the inferences from those trials; the inferences depend on unverifiable assumptions about the missing data. Ninety-four percent of trials with repeated follow-ups had reported late dropouts; more than one-half of them had >10% late dropouts and a quarter had >20%. Unlike early dropouts, the late dropouts can be included in an analysis effectively and make the MAR assumption more plausible, since interim measures of outcome are available on the late dropouts. Reasons for the dropouts are extremely important and should be collected, since they can be used to justify the analysis assumptions. Trials in this review frequently reported the reasons for the late dropouts, but mostly failed to justify the assumptions made to impute the missing data based on these reasons. Subjects who dropped out for different reasons were handled differently in the analysis in a few trials, but this approach still may not be appropriate. For example, in one trial (Baerwald et al., 2010) with a late dropout rate of 24% (191/810), reasons for dropouts included lack of efficacy (n=63), adverse events (n=55), withdrawn consent (n=39), violation of eligibility criteria (n=18), lost to follow-up (n=7), and other unspecified reasons (n=9). The primary analysis of reduction in pain score imputed missing values using either the WOCF if data were missing due to a withdrawal attributable to a treatment related adverse event, or the LOCF if the data were missing for any other reason. The implicit assumption of zero reduction in pain score from the point of discontinuation may not be appropriate in many situations. Subjects who discontinued due to lack of efficacy are more likely to return to their baseline score and BOCF may be more sensible. In the same way, subjects' withdrawal of consent is more likely to be associated with adverse events or difficulty in complying with study interventions (Gabriel & Mercado, 2011), and hence LOCF may not be appropriate in these situations.

3.6.6 Analysis strategy

In accordance with the ITT principle, most trials in the review analysed the data by the groups to which subjects were randomized regardless of the intervention received. However, many of these trials failed to measure outcome data on all randomized subjects and/or include all the subjects in the primary analysis. Consequently, only one-third of trials performed a FITT approach for the primary analysis. The proportion is comparable with that reported by Kruse et al. (2002) and Gravel et al. (2007); these studies examined the quality of RCTs published in general medical journals in 1999 and 2002, respectively. However, the proportion is higher than that reported by Baron et al. (2005). The lower proportion noted by Baron et al. (2005) may reflect changes in practice since 1994–2003, or may be due to their focus on trials with a particular outcome measure (structural outcomes in rheumatic diseases) without regard to whether or not it was a primary outcome. In many RCTs, secondary outcomes are less focused and less rigorously measured, analysed or reported (Hopewell et al., 2013), and this may explain the lower proportion of FITT analyses in Baron et al.'s (2005) study. An increasing number of guidelines (CONSORT 2001; CONSORT 2010; Food and Drug Administration, 2008) and endorsement of these guidelines by journals have improved the quality of reporting of RCTs (Turner et al., 2012), and hence may have improved the use of the ITT principle.

Early dropouts are one of the major challenges to performing a FITT analysis. The subset of the FITT sample (called the PITT sample) was the analysis choice in more than one-third of the reviewed trials with repeated follow-ups. It was clear from the CONSORT flow chart in the trial reports that none of the excluded subjects completed any post-baseline assessment; therefore, these trials were included in the PITT category in terms of early dropouts. It is worth noting that the descriptions on the PITT strategy (Table 3.5) did not include reasons for early dropout of subjects, but only indicated reasons for the exclusion from the analysis in many situations. For example, in table 3.5, the descriptions relating to

initiation of treatment, baseline assessment or post-baseline assessment did not provide reasons for the missing data in a trial – whether the missing outcome data resulted from withdrawals from the trial due to protocol violation or from loss to follow-up. Additionally, several previous studies have reported that PITT was inconsistently defined and the description did not reflect the actual analysis strategy followed (Baron et al., 2005; Gravel et al., 2007; Abraha & Montedori, 2010; Alshurafa et al., 2012). For example, in one trial in this review (Daniels et al., 2011) it was unclear why the authors defined the analysis sample based on a PITT description; all randomly assigned patients received at least one dose of study medication and had at least one post-baseline follow-up for the primary outcome (where no subject was excluded from the analysis) – indicating a FITT sample. Therefore the descriptions provided in trial reports need to be interpreted cautiously.

Late dropouts are also a challenge to a FITT analysis. A quarter of trials with late dropouts excluded them from the primary analysis (i.e. a CCA was adopted). In a pragmatic setting, exclusion of participants in a trial may limit interpretation of findings and therefore a FITT is recommended (Chan et al., 2013; Heritier et al., 2003), but it is clear that most trials in this review chose instead to exclude participants who had dropped out, whether late or early. Conceptually, if the principle of analysed-as-randomized is disturbed in any way, and for whatever reason, the chance of an imbalance in group comparison increases. At worst, early dropout may be an indication of selective 'exclusion'. Such exclusions of dropouts with no follow-up data may be reasonable if we can assume that the process of exclusion is protected against biased selection of an outcome or a data-driven preference for a particular analysis. Such protection is afforded when the criteria for exclusion of participants from the analysis are pre-specified in the protocol, and are not based on information related either to treatment allocation or to events or outcomes that occurred after randomization (Heritier et al., 2003). However, such exclusions should be limited in order to avoid selection bias (Heritier et al., 2003). Ideally, such decisions should also be made by a blind or independent observer.

3.6.7 Baseline comparison

Exclusion of randomized subjects can affect the baseline balance between arms achieved through randomization. Comparison of baseline variables between arms for those who were included in the analysis indicates how loss to analysis affects the randomization balance. Baseline-adjusted analyses are essential to control selection bias when imbalance exists; however, this may not remove bias due to the exclusions. In this review, it was noted that the baseline comparison was mostly performed on the FITT sample rather than the analysis sample for trials where some loss to analysis was reported. Twenty out of 34 and 11 out of 19 trials failed to report the baseline comparison based on the subjects included in the primary analysis when the loss to analysis was more than 5% and 10%, respectively. These trials may have failed to diagnose any imbalance between the intervention arms and hence not adjust for this potential bias in the analysis. For example, in a two-arm trial with 27% dropouts (original sample size = 158; 79 vs. 79), Fritsche et al. (2010) reported baseline comparison between two intervention arms for those initiated interventions (PITT sample, n=150; 79 vs. 71); the primary analysis did not adjust for potential contrast in baseline variables for those analysed (CCA sample, n=115; 60 vs. 55). This approach may sometimes be justified, as the proportion of dropouts here was similar across arms; however, an equal dropout rate between arms does not necessarily ensure a baseline balance between arms in a CCA sample as in an ITT sample. Their baseline comparison failed to observe any significant difference in baseline variables between the two study groups. The authors also reported a dropout analysis, in which comparisons of baseline variables between completers and dropouts were performed, but this was done without considering arm-wise stratification, which would need to be carried out to assess systematic imbalance in dropouts between study groups. The dropout analysis does not help to detect whether dropouts affect the randomization balance; however, it can be useful in assessing the plausibility of the MCAR assumption (Little, 1988). In this review, 10 trials reported a dropout analysis and most used this to

reasonably justify that dropouts did not affect the randomization balance. Many authors in this review highlighted equality in dropout rate between arms, possibly indicating a general view that equal dropout rate between arms would not lead to a biased estimate of treatment effect. However, bias is a function of both the frequency of and the reasons for the missing values in each arm (Wright & Sim, 2003).

3.6.8 Handling missing data

White et al. (2012) argue that including all randomized subjects in an analysis of an outcome with missing data is not enough; one should consider an appropriate method to handle the missing data. As mentioned in the background section, the validity of such an analysis depends on the correctness of assumptions made about missing data, which cannot be completely verified in most trials. Hence, trialists should attempt to justify the assumption based on the observed data. Although most of the reviewed trials reported the method used to handle the missing data, many of them failed to justify its adoption.

There are a few options available to account for early dropouts if baseline data are available, but few of them are generally valid (Wood et al., 2004) – particularly methods that are simple and involve a single imputation. Almost all of the trials reviewed excluded the early dropouts because of absence of an outcome assessment. The missing primary outcome values were imputed using BOCF in two trials and MI in one trial. MMRM was performed in one trial through considering the baseline as an outcome.

As regards late dropouts, these were improperly handled in many trials. A CCA includes only those subjects with complete data on the variables included in the analysis. In trials with repeated follow-ups, standard statistical methods, such as ANCOVA, exclude dropouts and does not consider availability of measurements at interim visits. These methods are likely to provide a biased result, unless the mechanism is MCAR, inefficient estimates, that is, estimates that have wide confidence intervals through lack of precision,

and loss of statistical power (Molenberghs & Kenward, 2007). More than a quarter of trials with repeated follow-ups excluded the late dropouts from the primary analysis. The proportion is lower than that reported by Wood et al (2004), in which nearly half of trials with longitudinal measurements used a CCA as the primary analysis. Additionally, the preference for a CCA was higher in trials with lower dropout rates compared to trials with a dropout rate greater than 20%. However, it is unclear whether the researchers erroneously believed that exclusions do not lead to a biased inefficient estimate of treatment effect in trials with lower dropout rates.

There are several single imputation strategies common in the practice of RCTs (Little & Rubin, 2002). Unlike CCA, single imputation methods do not exclude subjects with missing outcome assessment and aim to create a full dataset, which can then be analysed using the standard statistical methods by considering it as a real dataset. These methods are generally not recommended because they fail to account adequately for the uncertainty in the data and may produce biased estimates (Molenberghs & Kenward, 2007). Sixty percent (44/75) of trials with repeated follow-ups in the review used some sort of single imputation to replace the missing values. The findings of these trials are doubtful, as many of these methods are not valid even under MCAR (detailed in chapter 2).

In particular, the LOCF approach makes a very strong assumption, which is unlikely to be true, that the value of an outcome remains constant after dropout. This was the most frequently used imputation method in the review. In a retrospective analysis of randomized antidepressant efficacy trials published during 1965–2004, Woolley et al. (2009) found that the percentage of RCTs using the LOCF method had increased over time. Additionally, in a review of reports published in '*Arthritis and Rheumatism*' in 2005, Kim (2006) reported that LOCF was the most commonly used approach to handle missing data. The present review also yielded a similar finding: the approach was used in nearly a third of the trials with late dropouts, and was more prevalent in trials with a dropout rate

greater than 10%. The assumption of zero change after dropout is not justifiable in most trials, especially in trials comparing a new treatment against a standard treatment that is already proven effective. Given advances in methods that require less restrictive assumptions than LOCF and recommendations against using LOCF (Molenberghs & Kenward, 2007; Lane, 2008; National Research Council, 2010), the preference for this method is a major concern.

A substantial proportion of trials used imputations that require extreme assumptions; for example, the assumption that dropouts are 'non-responders' or have no change from baseline, without considering whether the subject was responding to treatment at the time of dropout. The review identified that a substantial proportion of trials used these kinds of imputations to replace the missing values. In 16% (12/75) of trials with late dropouts, the dropouts were simply classified as failure where the primary outcome was analysed on a dichotomous scale ("success or failure"). Additionally, in another four trials with late dropouts, the missing outcome data were replaced by baseline data where the primary outcome was analysed on a continuous scale. These imputations can sometimes provide useful bounds on the effect of the missing data, but this rarely produces an unbiased estimate of treatment effect. Additionally a few trials replaced missing values through a regression or linear extrapolation based imputation. Though these are based on the MAR assumption, use of a single imputation generally underestimates the standard error (Molenberghs & Kenward, 2007).

MI has recently received substantial attention in the literature as it helps to overcome the limitations of single imputations for handling incomplete data in RCTs (Sterne et al., 2009). The MI technique uses several stochastic imputations to incorporate the uncertainty surrounding missing values and it gives valid standard errors under MAR. A review of trials published during 2005–2009 in four general medicine journals (*BMJ*, *JAMA*, *Lancet* and *New England Journal of Medicine*) found a substantial increase in MI

use and reported that 9% of the trials published in 2009 used MI in some way (Mackinnon, 2010). Despite the evidence of increasing use of MI in the analysis of RCTs, only two trials in this review (Beaudreuil et al., 2011; Thorn et al., 2011) performed MI-based analyses and reported the results. Another three trials claimed to have done MI in some way as a sensitivity analysis, but no details or results were reported. Moreover, trials that performed MI failed to report the procedure adequately. Sterne et al. (2009) suggest guidelines for reporting analyses based on MI to prevent potential pitfalls with its application and to aid interpretation of its results. The guidelines specifically endorse the reporting (at least as online supplements) of details of the imputation modelling including: details of the software used; key settings for the imputation modelling; number of imputations performed; list of variables included in the imputation procedure; how the variables were handled; and importantly plausibility of MAR assumption based on the variables included in the imputation model.

Analysis methods that make use of available data on all time-points were limited to a small proportion of reviewed trials. FIML-based models can use all available longitudinal data without a need either to delete or to impute measurements. A FIML method such as MMRM is valid when the dropout mechanism is ignorable (Laird & Ware, 1982). Only 8% (6/75) of trials with late dropouts used MMRM to analyse the longitudinal outcome data without any imputation. Additionally, MMRM was performed after imputation of missing values using LOCF in two trials (Genevay et al., 2010; Fary et al., 2011), BOCF in one trial (Bliddal et al., 2011), and MI in another one (Thorn et al., 2011). The use of imputations such as LOCF and BOCF undermine the benefits of MMRM and the results may not be valid under MAR. Moreover, the use of MI prior to performing the MMRM is not necessary, as there is no obvious gain from doing so (Twisk et al., 2013).

Semi-parametric regression-based methods such as GEE can also use all available data. However, a standard GEE method is valid only under an MCAR assumption (Liang &

Zeger, 1986). In this study, two trials with late dropouts performed the standard GEE. Weighted GEE, in which weight is assigned at the subject level and calculated as the inverse of the probability for dropping out at the observed time of dropout, or MI-based GEE, in which MI is used before performing GEE, is preferred over a standard GEE because these methods can provide a valid estimate of treatment effect under MAR (Birhanu et al., 2011). No trials in this study presented results based on these methods.

3.6.9 Sensitivity analysis

Many researchers have agreed that analysis based on an MAR assumption is a reasonable starting point in many RCTs (Molenberghs & Kenward, 2007; National Research Council, 2010; White et al., 2011). Further, since there is no established set of tools to evaluate and distinguish one missing data mechanism from another, one should always be open to the possibility that the data are MNAR. Therefore, it is important to evaluate the sensitivity of the results and overall conclusions of a trial to possible departures from the MAR assumption by assuming a range of MNAR mechanisms. There is no established guideline or method in this matter as this is an active area of research (National Research Council, 2010). However, there are recommendations to explore the robustness of key inferences to possible departures from the expected missing data mechanism (Molenberghs & Kenward, 2007; National Research Council, 2010). The present study found that sensitivity analyses are infrequently and inappropriately used, and insufficiently reported. In the present era of the internet, authors have choices to publish sufficient details through online supplements if there is restriction in the main body of reports. It is unfortunate that many trials that performed a sensitivity analysis used the CCA or single imputation methods as the preferred approach in the sensitivity analysis. The use of CCA can be justified if additional exploration of the results under MAR and MNAR are performed. However, this was not the situation in these trials.

3.7 Limitations and generalizability

The review was restricted to trials published in a sample of high-impact factor speciality journals. The small number of journals evaluated may lower the generalization of the findings. However, the journals were from different publishers; the expectation is that journals from same publishers may have similar reviewing criteria. The selected journals were high-impact factor journals, which should bias the results towards a better methodology and reporting; hence, it is expected that the statistics reported here around appropriate use of methodology for ITT evaluation will be overstated rather than understated. Thus, the finding here that the minority of RCTs are performing a full ITT analysis (as per recommendation) is likely to be conservative; the application of this 'gold standard' evaluation method across all pragmatic trials published in MSCs may be less frequent (particularly in respect of publications within speciality journals in this clinical area). Appreciably, the largest priority publications in any clinical area are likely to be published in the highest impact generic journals such as *Lancet* and *BMJ*; these publications, though few in number, reflect the pinnacle of research in any area of clinical research and are likely to have greater methodological quality as they undergo more extensive reviewer scrutiny. Nonetheless, the focus here on speciality journals likely embraces the major portion of published RCTs in the area of MSCs.

3.8 Conclusion

It is important to take careful steps to minimize missing data in trial design and data collection. However, the occurrence of missing outcome data is not avoidable in the majority of trials. Thus, handling missing data is a major challenge when analysing trial data. Simple methods such as exclusion of subjects with missing data and LOCF were the frequently adopted approaches in handling the missing data, despite the availability of advanced methods and sophisticated software programmes to deal with the missing data more appropriately. Further, most of the trials failed to report a sensitivity analysis that

aimed to assess the impact on results and inferences if the assumption made on the missing data mechanism was wrong, i.e. by examining the robustness of the primary result to a range of plausible missing data assumptions. Since trials with a high proportion of missing data are highly sensitive to deviation from simple assumptions like MCAR, and the assumptions cannot be fully justified from the data, reporting of sensitivity analysis is crucial in these trials.

The systematic review raises concerns over the findings because of lack of / inappropriate use of the ITT principle, particularly in relation to missing data handling methods. The review highlights a need to conduct further evaluation and comparison of the frequently used methods such as CCA and LOCF with relatively efficient methods such as MMRM and MI. Of particular relevance to this PhD is to understand the merit of the different approaches in respect of key statistical parameters concerned with bias, precision and statistical testing. The goal of the following chapters of the thesis is to perform a comparative evaluation and present findings in a simple and clear way within a clinically meaningful context that makes it accessible to non-specialists without compromising the theoretical framework on which it is based.

Chapter 4: Simulation study: an overview of design

4.1 Introduction

As detailed in chapter 1, the primary objective of the thesis is to examine the relative performance of four missing data handling approaches – complete-case analysis (CCA), last observation carried forward (LOCF), mixed-effects model for repeated measures (MMRM) and multiple imputation (MI) methods – when analysing incomplete longitudinal clinical trial data with continuous outcome observations in relation to: spread of data, correlation between repeated measurements, trajectory pattern, sample size, dropout rate, and missing data mechanism. To meet the objective, four simulation studies were performed, and detailed below:

Study 1: This simulation study aimed to understand whether changes in data variability and correlation between repeated measurements affect overall accuracy of the missing data handling approaches.

Study 2: This simulation study aimed to understand whether changes in trajectory pattern over a study period and size of treatment effect at the endpoint affect the overall accuracy of the missing data handling approaches.

Study 3: This simulation study aimed to understand whether an increment in sample size in proportion to an expected dropout rate helps to achieve the required statistical power when using the missing data handling approaches.

Study 4: This simulation study aimed to compare two strategies for handling baseline data in MMRM analysis. These strategies are: (i) the baseline and post-randomization values are modelled as outcome variables and assume the baseline mean responses for the treatment groups are equal; and (ii) the baseline values are used as a covariate in the analysis of post-randomization

values, allowing different regression slopes. These strategies were detailed in chapter 2 (section 2.4.2).

This chapter presents detailed aspects of the design of these simulation studies.

4.2 Background

A Monte Carlo simulation – a numerical technique for conducting experiments on a computer – has a vital role in evaluating statistical methods. Within such a simulation study, a computer draws random samples with replacement from a population with known population parameters. The goal of any statistical method should be to make statistically valid inferences about population parameters from a random sample. A statistical method is recommended for practical use only if it provides results that are representative of the population parameters. Such recommendations are usually based on evaluation of statistical methods through theoretical proof and/or simulation studies (simulation being particularly relevant in cases where theoretical explanation is difficult or impossible to ascertain). Simulation studies can also be used to compare the performance of two or more procedures for addressing the same problem.

The evaluation of statistical methods using empirical data is limited by unknown population parameters: it is not possible to form valid judgements on missing values based on available empirical data. Hence, a legitimate comparison of missing data approaches cannot be performed using empirical data. In general, these simulation studies help to answer the following questions: (i) how do estimates of population parameters deviate from true values under various missing data scenarios? and (ii) how different are the results obtained when different missing data approaches are used?

A simulation study to evaluate missing data methods is performed by drawing multiple random samples of the same size from a realistic population with known parameters. Missing values are imposed on these random samples according to some plausible

missing data properties, such as missing data rate, pattern and mechanism. Parameters are estimated on these samples with imposed missing data, and these are compared with the true population parameters.

The design of each simulation study listed above varied somewhat depending on what was needed to address the question posed. A detailed description of the simulation procedure applied is provided below.

4.3 Simulation procedure

Simulations were performed using Stata version 12.0 (StataCorp, 2011). The syntax used for simulations are listed in appendix 3. A schematic diagram in figure 4.1 summarizes the main steps involved in the simulation methodology. The details are discussed below.

4.3.1 Step 1: Generating complete datasets

For simplicity, only the case of two-arm individually randomized clinical trials with equal number of subjects per arm was considered. Characteristics of the simulated data were similar to those of a (typical) large, phase 3, randomized clinical trial, where the primary outcome was measured at four time-points (baseline visit [visit 0] and three post-baseline visits [visit 1, visit 2 and visit 3]), and compared the effectiveness of two treatments (control group and experimental group) at the primary endpoint (visit 3), in the context of musculoskeletal conditions. The simulated outcome variable Y represents pain intensity, measured on a 0–100 numeric rating scale. A high score indicates greater severity of pain, and reduction of score over time therefore indicates an improvement in pain.

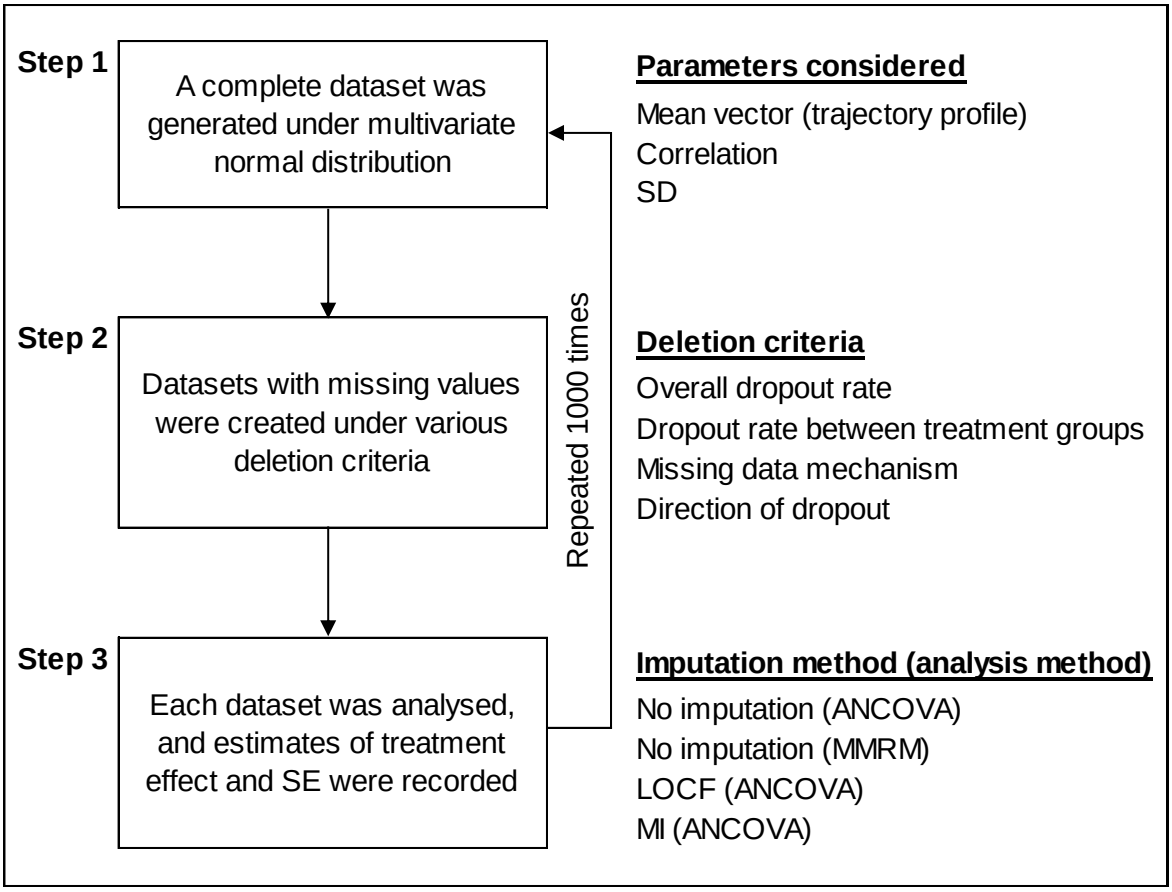


Figure 4.1: Schematic diagram to show the simulation procedures

In the simulation study, 1000 datasets (i.e. 1000 iterations) were generated under a multivariate normal assumption with a given mean vector and an unstructured covariance matrix. The properties of the selected mean vector and covariance structure are described below. The large number of iterations was used to obtain a robust estimate of the population parameters; the 1000 datasets obtained on a single trial scenario were meant to give an accurate view of the population distribution in respect of the specific trial scenario. Independence of the simulated datasets within a scenario was achieved by the use of a Monte Carlo simulation approach. Burton et al. (2006) point out that these generated datasets should also be completely independent for the different scenarios

considered. Different starting seeds were specified in the Stata simulation programme[5] to generate the independent complete datasets for each set of parameters considered with the multivariate normal distribution (Burton et al., 2006).

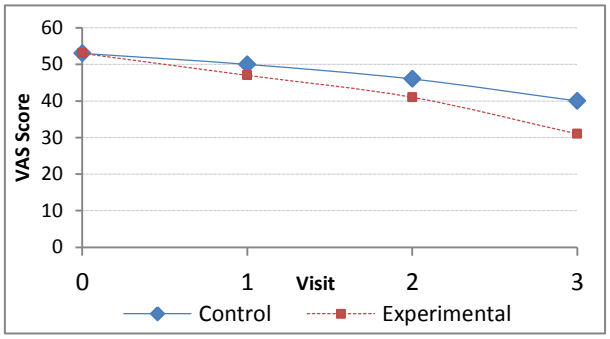
4.3.1.1 Choices of mean vector (means trajectory)

Four different sets of means trajectories with three treatment effect sizes relating to difference in mean pain score at the primary endpoint were assumed (Figure 4.2). Trajectory 1 assumed both treatments improved over time. However, the treatment group improved more and thus resulted in a treatment benefit over time (size of the treatment effect at the endpoint was assumed to be −9.0). This trajectory and the size of the treatment effect were chosen to mimic a typical (but hypothetical) scenario in trials of musculoskeletal disorders in primary care. This scenario was considered for all simulations unless otherwise specified.

[5] The Stata command used to simulate datasets is "simulate, rep(1000) seed(#): *command"*; where seed(#) sets the random-number seed.

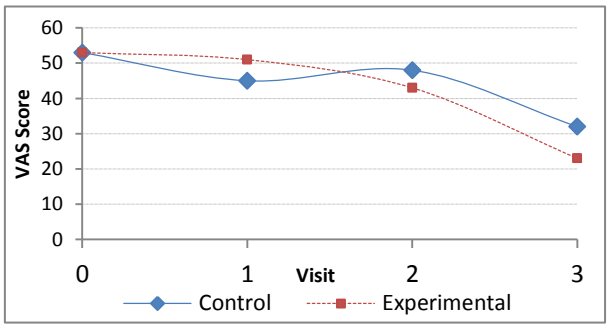
Trajectory 1 with treatment effect = −9.0

Visit	0	1	2	3
Control	53	50	46	40
Experimental	53	47	41	31



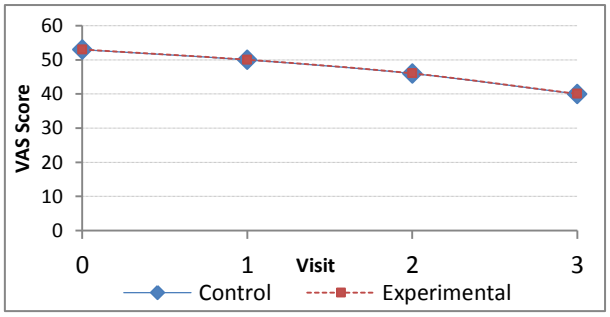
Trajectory 2 with treatment effect = −9.0

Visit	0	1	2	3
Control	53	45	48	32
Experimental	53	51	43	23



Trajectory 3 with treatment effect = 0.0

Visit	0	1	2	3
Control	53	50	46	40
Experimental	53	50	46	40



Trajectory 4 with treatment effect = −18.0

Visit	0	1	2	3
Control	53	50	53	48
Experimental	53	31	28	30

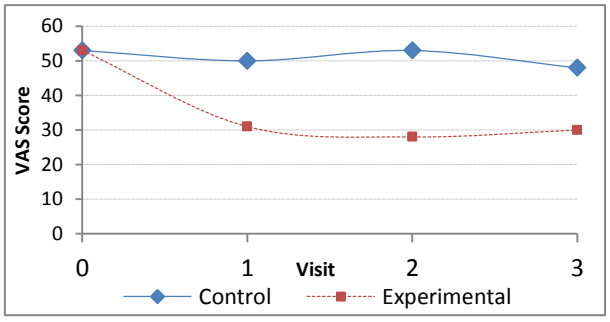


Figure 4.2: Assumed means trajectories

In addition, three more trajectory patterns – trajectories 2–4 – were also considered to address the objective of study 2 (i.e. to assess the impact of trajectory pattern and size of treatment effect on inferences from the missing data handling approaches). Trajectory 2 assumed a similar treatment effect (−9.0) at the primary endpoint as in trajectory 1; however, the control group showed better improvement during the initial visit (visit 1) than the experimental group. Trajectory 3 assumed both treatments improved equally well, and reflected the null hypothesis that there was no difference between treatments at the primary endpoint (i.e. size of the treatment effect at this endpoint is zero). Trajectory 4 assumed the experimental group improved quickly but then showed little change over time; however, there was minor improvement in the control group and a treatment effect of −18.0 at the primary endpoint was assumed in favour of the experimental group.

4.3.1.2 Choices of unstructured covariance matrix

The extent of correlation between repeated measurements may vary widely across different RCTs. Similarly, the degree of data variability is likely to differ considerably across studies. Such differences are likely to arise from a variety of factors such as the type of outcome being measured. Hence, it is important to consider a range of covariance matrices to reflect the possible variation in data across studies. I used six covariance matrices – one in each scenario: weak correlation with low SD (WL), moderate SD (WM) and high SD (WH); strong correlation with low SD (SL), moderate SD (SM) and high SD (SH). The correlation and SD matrices are listed in table 4.1. The defined variance-covariance matrices were based on, and similar in structure to, reported publications that were included in the systematic review in chapter 3. In keeping with the variance-covariance patterns observed in the systematic review, the assumed covariance matrices exhibit parameters that reflect the variance of the outcome variable increasing across time and the correlation diminishing as the time-lag increases.

Table 4.1: Correlation and SD matrices for simulation scenarios

Correlation matrices

1. Weak

	t0	t1	t2	t3
t0	1			
t1	0.45	1		
t2	0.39	0.41	1	
t3	0.30	0.34	0.40	1

2. Strong

	t0	t1	t2	t3
t0	1			
t1	0.75	1		
t2	0.63	0.71	1	
t3	0.54	0.59	0.66	1

Standard deviation (SD) matrices

	1. Low	2. Moderate	3. High
t0	10.2	14.1	24.5
t1	10.7	14.6	25.0
t2	11.4	16.9	25.7
t3	12.2	17.7	26.5

4.3.1.3 Sample size calculation

In general, sample size calculation was carried out based on an ANCOVA model – which may considerably reduce the number of subjects required for an RCT in comparison with the calculation based on *t*-test. The extent of reduction in the overall sample size is directly associated with the amount of correlation between baseline and endpoint data (Borm et al., 2007). Given the true treatment effect of −9.0 at the endpoint and various covariance patterns, the number of subjects per group required to detect the true difference with 90% power and 5% type 1 error rate is given in table 4.2. However, for simulations in study 2 – effect of trajectory patterns and the size of treatment effect – a fixed sample size of 60 per group across the trajectories was used in simulations.

Table 4.2: Calculated sample size under various covariance patterns

True treatment effect	Correlation between baseline and endpoint	SD at endpoint	Sample size per group
−9.0	Weak	Low	37
		Moderate	75
		High	168
	Strong	Low	30
		Moderate	60
		High	130

One commonly used approach to account for loss of power due to dropouts in hypothesis tests or confidence intervals is to inflate the sample size in proportion to the anticipated dropout rate. Hence, to evaluate the effect of the ad hoc approach used for the sample size calculation, simulation study 3 was performed using trajectory 1 and the WM covariance matrix with sample size provided in table 4.3. In this simulation study, datasets with a sample size not inflated for expected dropout rate – as used with other simulation studies – were compared against datasets with a sample size inflated for dropouts. The study was replicated under four scenarios: combinations of dropout rates (10% and 30%) and desired statistical power (80% and 90%) in the absence of dropouts. This simulation study is intended to illustrate how an increase in sample size protects against the loss of power due to dropouts.

Table 4.3: Sample size used for study 3 – effect of sample size

#	Dropout rate	Inflated for dropouts	Sample size per group[1]	
			80% power in the absence of dropouts	90% power in the absence of dropouts
1	10%	No	57	75
2		Yes	63	84
3	30%	No	57	75
4		Yes	81	108

[1]For a given WM covariance matrix

4.3.2 Step 2: Generating missing data

From the complete datasets, incomplete datasets were created by using some pre-specified deletion criteria. In all datasets, baseline measurements were complete for all subjects. In most longitudinal RCTs, the majority of missing data are caused by subjects discontinuing the trial prior to the primary endpoint. In such cases, the resulting missing data have a monotone pattern, meaning that once a subject has a missing observation for some visit, data will be missing for all subsequent visits. Typically, however, in practice there will be some small amount of non-monotone missing data (when subjects skip intermediate visits but returns for evaluations at subsequent visits). However, in an endpoint analysis, the major concern is whether the outcome has been observed at the primary endpoint. Hence, for simplicity, only a monotone missing data pattern was considered for the simulation studies. The monotone missing data pattern was imposed according to predefined dropout rates at each visit and in respect of a variety of assumed dropout mechanisms.

4.3.2.1 Choices of dropout rates

I considered a series of overall dropout rates with equal and unequal dropout rate between study groups. The selection was generally based on findings of the systematic review in chapter 3; the review identified that around 60% of studies had more than 10% and 15% had more than 30% dropouts at their primary endpoint. The percentages of missing data imposed on the simulation studies are summarized in table 4.4. The baseline (visit 0) measurement was assumed to be always observable; however, situations where some subjects had no post-baseline data were considered. All the studies were mainly centred on 10% and 30% dropout rate; however, MNAR situations were further explored with 20%, 40%, and 50% dropout rates. The use of a range of dropout rates – low to high – helps to assess the impact that level of dropout rates, whether equal or unequal between study groups, has on estimation of treatment effect.

Table 4.4: Planned cumulative dropout rate (%)

Dropout rate	Group	Equal dropout rates (EQ)				Higher dropout in experimental group (HE)				Higher dropout in control group (HC)			
		t0	t1	t2	t3	t0	t1	t2	t3	t0	t1	t2	t3
10%	Control	0	2	6	10	0	2	4	6	0	4	8	14
	Experimental	0	2	6	10	0	4	8	14	0	2	4	6
20%	Control	0	4	10	20	0	4	8	14	0	8	16	26
	Experimental	0	4	10	20	0	8	16	26	0	4	8	14
30%	Control	0	10	20	30	0	5	15	20	0	10	25	40
	Experimental	0	10	20	30	0	10	25	40	0	5	15	20
40%	Control	0	10	20	40	0	10	20	30	0	15	30	50
	Experimental	0	10	20	40	0	15	30	50	0	10	20	30
50%	Control	0	10	25	50	0	10	25	40	0	15	30	60
	Experimental	0	10	25	50	0	15	30	60	0	10	25	40

4.3.2.2 Dropout mechanism

Dropout rates, as listed in table 4.4, were imposed based on various dropout mechanisms: missing completely at random (MCAR); missing at random dependent on baseline data (MAR-B; where an observation that triggered a dropout was a baseline observation) or last observed outcome value (MAR-L; where an observation that triggered a dropout was a last observed value); and missing not at random (MNAR; where an observation that triggered a dropout was the observation itself). The MAR and MNAR dropout mechanisms were implemented under two quite contrasting situations to fully understand the robustness of the results under the considered missing data handling approaches. The observations at a time-point that triggered dropouts were a random selection within upper or lower q^{th} percentile (where q = dropout rate at the endpoint plus

20%) of the observations at that time-point, depending on the following situations[6]. In the first situation, the dropouts were constrained to be in the ***same direction*** in both of the treatment groups whereby the same reasons for dropout generally applied across the groups. In this situation, subjects who had performed poorly (subjects with high values) were randomly dropped from both the control and the experimental group. In the second situation, the dropouts were in ***opposite directions*** across the groups in the sense that the reason for dropout was allowed to differ between the groups. In this situation, subjects who had performed poorly in the control group and well in the experimental group were randomly dropped. Admittedly, these situations are deliberately quite extreme since the interest is in evaluating possible large-scale impact and deviation in treatment response through differential and skewed dropout mechanisms. These situations were chosen to provide a more telling picture of what happens under various missing data situations, since one cannot predict what would have been the situation in a real incomplete dataset. Detailed descriptions of the dropout mechanisms are given below.

4.3.2.2.1 MCAR

In the MCAR dataset, all missing outcome values were selected randomly from the complete dataset.

4.3.2.2.2 MAR-B1

In the first MAR dataset (MAR-B1), absences at different time-points were selected in each group randomly from a subpopulation for which baseline values were higher than p^{th} percentile of the baseline variable; where p is equal to: 100 – (dropout rate + 20%). In this

[6] A similar strategy of implementing dropout rates based on a cut-off value was reported in a few publications (Hedeker & Gibbons, 2006; Enders & Gottschall, 2011)

dropout mechanism, a random sub-sample of subjects with higher baseline values was dropped from both the control and experimental groups.

4.3.2.2.3 MAR-B2

In the second MAR dataset (MAR-B2), absences at different time-points were selected randomly from a subpopulation for which baseline values were higher than the p^{th} percentile of the baseline variable in the control group and lower than the q^{th} percentile of the baseline variable in the experimental group; where q is equal to dropout rate plus 20%. In this dropout mechanism, a random sub-sample of subjects with high baseline values from the control group and another random sub-sample of subjects with low baseline values from the experimental group were dropped.

4.3.2.2.4 MAR-L1

In the third MAR dataset (MAR-L1), absences at different time-points (t) were selected in each group randomly from a subpopulation for which values of the outcome variable Y measured at time t-1 were higher than the p^{th} percentile of Y at time $t-1$. In this dropout mechanism, a random sub-sample of subjects with high last observed outcome values were dropped out from both the control and the experimental groups.

4.3.2.2.5 MAR-L2

In the fourth MAR dataset (MAR-L2), absences at different time-points were selected randomly from a subpopulation for which values of the outcome variable Y measured at time $t-1$ were higher than the p^{th} percentile of Y at time $t-1$ in the control group and lower than the q^{th} percentile of Y at time $t-1$ in the experimental group. In this dropout mechanism, subjects with high last observed outcome values were randomly dropped from the control group and subjects with the low values from the experimental group.

4.3.2.2.6 MNAR-1

Regarding MNAR, two datasets were created based on deletion restrictions. In the first MNAR dataset (MNAR-1), absences at different time-points were selected in each group randomly from a subpopulation for which values of the outcome variable Y measured at time t were higher than the p^{th} percentile of Y at that time-point. Here the eliminated observations represent a random sub-sample of high values for the outcome variable at that particular time-point in both the control and experimental groups.

4.3.2.2.7 MNAR-2

In the second MNAR dataset (MNAR-2), observations were eliminated for subjects with high values in the control group and subjects with low values in the experimental group. In this dropout mechanism, absences at different time-points were selected randomly from a subpopulation for which values of the outcome variable Y measured at a time t was higher than the p^{th} percentile of Y at that time-point in the control group and lower than the q^{th} percentile of Y at that time-point in the experimental group.

4.3.3 Step 3: Imputation and analysis methods

As detailed in chapter 1, the objective of the thesis is to compare the missing outcome data handling approaches in longitudinal trials, and the following approaches were considered:

i. CCA, where no imputation was performed and the data were analysed using ANCOVA[7] with the outcome at the endpoint as a response variable and baseline

[7] The Stata command used to estimate the treatment effect at the endpoint is: *regress y3 i.group baseline, where y3 is the outcome at the primary endpoint.*

score as a covariate. This method excludes subjects with missing data (i.e. listwise deletion).

ii. MMRM[8] with baseline score as a covariate, in which no imputation was performed. The baseline-as-covariate model included baseline-by-time interactions to allow for different regression slopes. Kenward et al. (2010) recommend restricted maximum likelihood (REML) estimation in conjunction with the MMRM model. The within-subject error was modelled using an unstructured variance-covariance pattern.

iii. LOCF-based ANCOVA analysis, where missing outcome values were replaced by the last observed value and then analysed using an ANCOVA model.

iv. Multiple imputation, as implemented in Stata (mi command)[9], was used to create multiple, say m, complete datasets from the simulated missing data. These m complete datasets were analysed independently using an ANCOVA model. Estimates of parameters of interest were averaged across the m copies to give a single estimate. Standard errors were computed according to the "Rubin rules" (Rubin, 1987) to incorporate additional uncertainty due to the missing data. These steps were performed using the Stata mi estimate[10] command. As a rule of thumb, Bodner et al. (2008) and White et al. (2011b) suggested the number of

---

[8] The Stata command used to estimate the treatment effect at the endpoint is: *xtmixed y i.time##c.baseline i.rand1 i.rand2 i.rand3 if time>0* || *id:, reml nocons res(uns, t(time))*; where rand1, rand2, and rand3 are the interaction terms defined as (group==1)*(time==1), (group==1)*(time==2), and (group==1)*(time==3) respectively. The term 'reml' indicates the usage of restricted maximum likelihood method for the estimation of parameters; the term 'nocons' used to suppress random intercept term from the random-effects equation; and the term 'uns' used to specify the unstructured covariance structure.

[9] *mi impute monotone (reg) y1 y2 y3 = baseline group, add(m)*; where m is the number of imputations.

[10] *mi estimate: reg y3 i.group baseline*

imputations, m, should parallel the percentage of subjects with missing data. For example, with 30% dropout rate, m should be 30.

In the study 4, I compared a variant of MMRM (by considering the baseline as an outcome[11], in which baseline data was included as part of the outcome variable and the baseline mean responses for the treatment groups were assumed to be equal) with the approach given in (ii), whereby baseline data was considered a covariate. Both models were evaluated using the Stata *xtmixed* command where adjustment for small samples [e.g. Kenward-Roger correction (Kenward & Roger, 1997) in SAS *proc mixed*] is not implemented. However, a comparison of these models with and without the Kenwrad-Roger correction using SAS *proc mixed* was performed.

4.4 Measures of performance

Quantities used to assess the performance of various missing data strategies were bias, accuracy, coverage probability, width of confidence interval, and statistical power. Details of these measures are presented below.

4.4.1 Bias

Bias is defined as the difference between the average value of estimated treatment effects over the simulation repetitions and the true parameter for treatment effect (i.e. raw bias = true parameter − average estimate of the parameter). Negative bias indicates underestimation of treatment effect and positive bias indicates overestimation of treatment effect. The raw bias, percentage bias (bias as a percentage of the true parameter) and standardized bias (bias as a percentage of SD of the parameter) are recommended assessments of bias (Burton et al., 2006). However, the raw bias was reported because

[11] *xtmixed y i.time i.rand1 i.rand2 i.rand3 || id:, reml nocons res(uns, t(time))*

the main objectives of the thesis includes the effect of size of the true parameter (study 2), where percentage bias would not be an appropriate measure, and the effect of size of the uncertainty in parameter estimates (study 1), where standardized bias would not be an appropriate measure of bias.

4.4.2 Overall accuracy of the estimate

Bias and variance of an estimate are often combined into a single measure called mean squared error (MSE) as a measure of overall accuracy of the estimate (Burton et al., 2006). It is the average of squared difference between the estimated treatment effects ($\hat{\beta}$) and the true parameter (β) over repeated samples. MSE is equal to the sum of the variance and the squared bias of the estimated treatment effect. For easier interpretation, the square root of the MSE – root-mean-square error (RMSE) – is reported, to place it on the same scale as the parameter (Collins et al., 2001).

4.4.3 Coverage of confidence interval

The actual coverage of the nominal 95% CI of the estimated treatment effect is the proportion of time that nominal intervals contain the true treatment effect across all simulation replications. Since the 95% CI aims to contain the true treatment effect with a probability of 0.95, actual coverage should be approximately equal to the nominal coverage of 95% if the missing data strategy works well. A coverage larger than 95% indicates too wide (imprecise) CIs whereas a coverage smaller than 95% indicates too narrow (too precise) CIs. Over-coverage suggests the results are too conservative, as more simulations will fail to find a significant result when there is a true treatment effect, thus leading to loss of statistical power with too many type II errors (Burton et al., 2006). In contrast, under-coverage is unacceptable as it indicates over-confidence in the estimates, since more simulations will incorrectly detect a significant result, which leads to larger than expected type I errors (Burton et al., 2006). Burton et al. (2006) suggest that an observed

coverage falling inside the interval 93.6%—96.4%[12] is considered to be an acceptable coverage of the nominal 95% CI. If the coverage is accurate, the probability of type I error (wrongly rejecting a true null hypothesis) will also be accurate (Burton et al., 2006; Collins et al., 2001).

4.4.4 Average width of confidence interval

Subject to correct coverage, a confidence interval of an estimate of a treatment effect should be narrow, because a shorter interval will reduce the probability of type II error (failure to accept a true alternative hypothesis). Therefore, if one method has a similar or higher coverage probability than another, but yields substantially narrower CIs, then it should be preferred (Collins et al., 2001).

Average width of CI $= \sum_{i=1}^{B} \frac{2*t_{(1-\frac{\alpha}{2},df)} SE(\widehat{\beta_i})}{B}$, B *is the number of repetitions*[13]

4.4.5 Statistical power

The empirical power is calculated as the proportion of simulation samples in which the null hypothesis of no effect is rejected at the 5% two-tailed nominal significance level when the alternative hypothesis is true (Burton et al., 2006). As noted in section 4.3.1.3, simulations under study 1 were planned to have a 90% nominal statistical power to detect the true treatment effect in the absence of missing data, while study 3 aimed to determine the size of the sample required to achieve 90% empirical power in the presence of missing data. Note that study 2, which aimed to understand the effect of the size of the treatment effect

[12] SE of the nominal coverage probability (p), $SE_{(p)}=\sqrt{p(1-p)/B}$; where B is the number of repetitions.

[13] *xtmixed* command in Stata does not compute df; it assumes z-distribution instead of t-distribution.

and trajectory pattern, does not report the empirical power since the scenarios under this study were a mixture of true null and true alternative hypotheses.

4.5 Summary of simulation scenarios

Under study 1, 252 scenarios were evaluated for a given mean vector, as summarized in table 4.5:

Table 4.5: Simulation scenarios under study 1

SD	Correlation	Dropout rate*	Dropout rate between groups	Missing data mechanism	Direction of dropouts[#]
• Weak • Moderate • High	• Weak • High	• 10% • 30%	• Equal between groups (EQ) • High in experimental group (HE) • High in control group (HC)	• MCAR • MAR-B[†] • MAR-L[‡] • MNAR	• Same direction • Opposite direction

*Additional explorations with 20%, 40%, and 50% dropout rates under MNAR were performed; [†]MAR dependent on baseline value; [‡]MAR dependent on last observed value; [#]not valid for MCAR.

Under study 2, 84 scenarios were evaluated for a given moderate SD and strong correlation, as summarized in table 4.6:

Table 4.6: Simulation scenarios under study 2

Mean vector	Dropout rate	Dropout rate between groups	Missing data mechanism	Direction of dropouts*
• Trajectory 1 with true treatment effect of −9.0 at the endpoint • Trajectory 2 with true treatment effect of −9.0 at the endpoint • Trajectory 3 with true null effect at the endpoint • Trajectory 4 with true treatment effect of −18.0 at the endpoint	• 30%	• Equal between groups (EQ) • High in experimental group (HE) • High in control group (HC)	• MCAR • MAR-B† • MAR-L‡ • MNAR	• Same direction • Opposite direction

†MAR dependent on baseline value; ‡MAR dependent on last observed value; *not valid for MCAR.

Under study 3, 168 scenarios were evaluated for a given mean vector, moderate SD and weak correlation, as summarized in table 4.7:

Table 4.7: Simulation scenarios under study 3

Sample size	Desired power	Dropout rate	Dropout rate between groups	Missing data mechanism	Direction of dropouts*
• Not adjusted for dropout • Adjusted for dropout	• 80% • 90%	• 10% • 30%	• Equal between groups (EQ) • High in experimental group (HE) • High in control group (HC)	• MCAR • MAR-B† • MAR-L‡ • MNAR	• Same direction • Opposite direction

†MAR dependent on baseline value; ‡MAR dependent on last observed value; *not valid for MCAR.

Under study 4, 42 scenarios were evaluated for a given mean vector, moderate SD and strong correlation, as summarized in table 4.8:

Table 4.8: Simulation scenarios under study 4

Baseline handling	Dropout rate	Dropout rate between groups	Missing data mechanism	Direction of dropouts*
• Baseline as part of an outcome vector • Baseline as a covariate	• 30%	• Equal between groups (EQ) • High in experimental group (HE) • High in control group (HC)	• MCAR • MAR-B† • MAR-L‡ • MNAR	• Same direction • Opposite direction

†MAR dependent on baseline value; ‡MAR dependent on last observed value; *not valid for MCAR.

4.6 Discussion and conclusion

To address the objectives of the thesis, four simulation studies were planned, and 546 scenarios were explored. Among the four simulation studies, the first three were focused on the comparison of performance of CCA, LOCF, MMRM and MI in relation to missingness properties (i.e. overall dropout rate, dropout rate between groups, dropout mechanism, and direction of dropouts) along with complete data characteristics (i.e. correlation between repeated measurements and data variability, means trajectory, and sample size). The remaining simulation was used to assess the robustness of estimates in respect of different approaches to analysis in MMRM. Little and Rubin (2002) point out that the degree of bias and imprecision depends on the extent to which complete and incomplete cases differ and on the parameters of interest, in addition to the proportion of missing data and the type of missing data mechanism. As noted in the literature review (Chapter 2, section 2.4.2), most previously reported simulation studies that compared missing data approaches did not explore the impact of factors other than the proportion of missing data and the missing data mechanism. In the present simulation study, I additionally considered the effect of deviation in mean trajectory, data variability, correlation between repeated measurements, dropout rate between groups (equal vs.

unequal), and direction of dropouts. In fact, apart from information on sample size, the overall dropout rate and dropout rate between groups, other information on the data (actual estimates of correlation, SD, and mean; dropout mechanism; direction of dropouts) are unknown when some missing data is present. However, the present simulation provides an expected range of deviation around the true parameters, nominal coverage probability and nominal power under the studied missing data approaches.

For the purpose of these simulation studies, focus is limited to analysis of a single continuous variable measured over time with monotone missingness. In RCTs, outcome data are usually measured at more than one follow-up; however, the aim of these trials is usually limited to comparing the effect of two or more treatments at a specific time-point (Verbeke & Molenberghs, 2005). Unless it is important to know how study participants have reached the study endpoint, a simple comparison of the treatment groups at the primary endpoint is often recommended and adequate to demonstrate the treatment effect, if any (European Medicines Agency, 2006; Verbeke & Molenberghs, 2005). As noted in the systematic review, nearly 90% of trials measured outcome in a longitudinal fashion; all of them limited the assessment of the treatment effect to the primary endpoint. Therefore, the simulation study was restricted to the endpoint analysis and did not assess the growth factor.

The trajectory profiles, which were considered in this simulation, represent a situation where reduction in baseline score indicates an improvement; hence, a negative sign in treatment effect indicates that reduction in score was greater in the experimental group than in the control group. The reason for this choice is that most outcomes in musculoskeletal trials (and therefore of interest within the Arthritis Research UK Centre) are such that a treatment effect implies decrease in baseline score. For example, pain scales, depression/anxiety scales, disability scales are usually scored so that low scores (often 0) denote no pain, no depression/anxiety, or no disability, whereas the highest

scores denote maximum pain, highest depression/anxiety or most severe disability. Hence, patients present with high scores (denoting increased severity of the condition) and the goal of the treatment is to lower the score. In bias estimation of treatment effect, based on the trajectory patterns considered, a negative bias therefore denotes an underestimation of the treatment effect and a positive bias denotes an overestimation of the treatment effect.

A range of summary statistics were considered for the simulation parameters – means, SDs, and correlation between repeated measurements – to understand the impact of these parameters (in study 1 and 2) on bias and loss of precision under the studied missing data approaches. In particular, study 1 evaluates the impact of SD and correlation on bias and loss of precision in the presence of missing data. A 'moderate SD' was chosen in order to ensure a moderate effect size (~0.5 with a treatment effect of −9.0) at the primary endpoint; thereafter, a weak and a high SD were defined in relation to the moderate SD. Additionally, two plausible correlation structures for association between repeated measurements (strong and weak) were considered – and have been used in a previous study by Siddiqui et al. (2009). Study 2 was planned in order to evaluate the effect of mean trajectories with the same treatment effect at the primary endpoint (trajectories 1 and 2 with treatment effect −9.0) and the effect of mean trajectories with different treatment effect at the endpoint (trajectories 1, 3 and 4 with treatment effects −9.0, 0 [no effect] and −18.0 respectively).

In practice, there is no way to ascertain missing values with certainty. Thus, it is necessary to make assumptions, which are often unverifiable in an incomplete data, about the missingness. It is crucial, therefore, to assess the performance of methods to deal with missing data in relation to a variety of scenarios – especially under quite contrasting scenarios – in order to understand the robustness of the results under the missing data handling approaches to the extreme situations and to aid interpretation of findings from an

incomplete dataset. In this simulation study, monotone missing data were generated with pre-specified dropout rates with equal and unequal dropout rate between groups – one with a higher dropout rate in the experimental group and another with a higher dropout rate in the control group – under the four missing data mechanisms: MCAR, MAR-B, MAR-L and MNAR. The difference between complete and incomplete data becomes substantial when the missing data mechanism changes from MCAR towards MNAR. Further, the MAR and MNAR dropout mechanisms were implemented under two quite contrasting scenarios. In the first scenario, dropouts were in the same direction in both study groups – dropouts were a random sub-sample of subjects who did poorly at baseline, at the last follow-up and the time at which they were dropped out under MAR-B1, MAR-L1 and MNAR-1 respectively. In the second scenario, dropouts were in opposite directions between study groups – dropouts were a random sub-sample of subjects who did poorly in the control group and those who did well in the experimental group at baseline, at the last follow-up and the time at which they were dropped out under MAR-B2, MAR-L2 and MNAR-2 respectively. Further scenarios contrasting to the scenarios listed above – (i) dropouts were subjects who did well in both study groups; (ii) subjects who did well in the control group and those who did poorly in the experimental group – were also explored and detailed in appendix 4 (Tables 1 and 2). The dropout scenarios were constructed so as to evaluate the direction and magnitude of the performance indicators – bias, coverage and power – of the missing data handling methods considered.

In an endpoint analysis, an important consideration is whether an outcome was measured at baseline and the primary endpoint. As detailed in chapter 2, a standard ANCOVA does not take into account the intermediate outcome measurements, and an MMRM model (to estimate the treatment effect at the endpoint) is less likely to be influenced by the intermediate outcome when the number of intermediate outcome assessments is small and when the outcome is measured at the endpoint. Simulated datasets with intermittent

missing values could have been used without altering the main point because either a standard ANCOVA (with no imputation of missing values), LOCF imputation, MI, or an MMRM model (with no imputation of missing values) was specified in order to assess treatment effect at the primary endpoint. Further, the motivation for the dropout setting is that monotone patterns are easier to address compared to arbitrary patterns of missing data; the computations needed to handle them are less cumbersome, and mechanisms producing them are easier to conceptualize and model. Dropout is often the dominant type of missingness in longitudinal studies, which partially explains why so many articles concerning dropout have recently appeared in statistics and biostatistics journals.

The relative performance of the methods – CCA, MMRM, LOCF and MI – was assessed in each scenario using performance indicators: (a) raw bias in estimate of treatment effect; (b) root-mean-square error; (c) coverage probability of 95% CI; (d) average width of the 95% CI; and (e) statistical power. Study 1 results are presented in chapter 5; the results of the remaining studies (study 2–4) are presented in chapter 6.

Chapter 5: Simulation study - findings 1

5.1 Introduction

This chapter presents the results of the simulation study 1 (detailed in chapter 4) that was designed to examine the relative performance of four missing data handling approaches – complete-case analysis (CCA), last observation carried forward (LOCF), mixed-effects model for repeated measures (MMRM) and multiple imputation (MI) methods – when analysing incomplete longitudinal randomized clinical trial (RCT) data under four missing data mechanisms: missing completely at random (MCAR), missingness dependent on baseline value (MAR-B), missingness dependent on last observed value (MAR-L), and missing not at random (MNAR). Analysis of covariance (ANCOVA) was used in conjunction with CCA, LOCF and MI; a baseline-as-covariate model was specified with MMRM. The comparison was in respect of various levels of experimental conditions typical of an RCT: levels of data variability, levels of correlation between repeated assessments, levels of dropouts, equal and unequal dropout rate between groups, and missing data mechanisms. As stated in the introduction chapter, the goal of this simulation study is to answer the following research questions:

i. Within and across the missing data handling approaches, does the accuracy (in terms of bias and precision) of the estimate of treatment effect at the primary endpoint vary by data variability, correlation between repeated assessments, and proportion of dropouts between groups, in addition to overall dropout rate and missing data mechanism?

ii. Within and across the missing data handling approaches, does the coverage and width of the confidence interval for the estimate of treatment effect at the primary

endpoint vary by data variability, correlation between repeated assessment, and proportion of dropouts between groups, in addition to overall dropout rate and missing data mechanism?

iii. Within and across the missing data handling approaches, does the statistical power to detect the true treatment effect at the primary endpoint vary by the spread of data, correlation between repeated assessment, and proportion of dropouts between groups, in addition to overall dropout rate and missing data mechanism?

It should be noted that simulation results, which are presented in this chapter, were based on a trajectory profile (trajectory 1; section 4.3.1.1 in chapter 4) with a treatment effect of −9.0 at the primary endpoint (i.e. at fourth visit). The effect of levels of trajectory profile and levels of treatment effect will be presented in chapter 6. The trajectory profiles that were considered in this thesis represent a situation where reduction in baseline score indicates an improvement (e.g. pain score); hence, a negative sign in treatment effect indicates that reduction in score was more in the experimental group than in the control group.

To understand the effect of data variability and correlation between repeated assessments on the relative performance of the missing data handling approaches, six unstructured variance-covariance matrices (WL, WM, WH, SL, SM, and SH) were considered, as detailed in chapter 4, with combinations of three SD matrices – low (L), moderate (M) and high (H) – and two correlation matrices – weak (W) and strong (S). Although the selection of the moderate SD ensured an effect size of approximately 0.5 at the endpoint (with a treatment effect of −9.0), the classification of SD matrices was purely based on the magnitude of the SDs. However, the classification of correlation matrices was based on whether the magnitude of correlation was close to zero (i.e. no correlation) or one (i.e. perfect correlation). Sample sizes were calculated in order to ensure 90% nominal power in all considered scenarios when there is no

missing data. Throughout, results are presented on all six levels of the variance-covariance structures.

As detailed in chapter 4, monotone missing data were generated with pre-specified dropout rates with equal and unequal dropout rate between groups – higher in the experimental group or higher in the control group – under the four missing data mechanisms: MCAR, MAR-B, MAR-L and MNAR. The MAR and MNAR dropout mechanisms were implemented under two quite different situations to help ascertain the robustness of the results to extremely contrasting situations. In the first situation, dropouts were in the same direction in both groups (the corresponding missing data mechanisms are denoted as MAR-B1, MAR-L1 or MNAR-1). In this situation, the value of an observation at a time-point that triggered dropout is higher than a threshold value corresponding to the p^{th} percentile of observations at that time-point in both control and experimental group; where $p = 100\% - (dropout\ rate + 20\%)$. In the second situation, dropouts were in opposite directions between the groups (the corresponding missing data mechanisms are denoted as MAR-B2, MAR-L2 or MNAR-2). In this situation, the value of an observation at a time-point that triggered dropout is higher than a threshold value corresponding to the p^{th} percentile of observations at that time-point in the control group and lower than an another threshold value corresponding to the q^{th} percentile of observations at that time-point in the experimental group; where $q = dropout\ rate + 20\%$.

In the following sections, I present the simulation results to address the accuracy, CI coverage and width, and statistical power of CCA, LOCF, MMRM and MI under various scenarios within each missing data mechanism. Section 5.2 presents the accuracy of the missing data handling methods in respect of research question 5.1-i, section 5.3 presents the CI coverage and width of the methods in respect of research question 5.1-ii, and section 5.4 presents the empirical power of the methods in respect of research question 5.1-iii. The

results are presented in graphs, and the corresponding tables are listed in appendix 5 (Tables 3–22).

5.2 Bias and precision

In this section, the simulation results are presented in terms of raw bias (i.e. true treatment effect minus observed treatment effect) and RMSE under various scenarios within each missing data mechanism. A negative bias indicates an underestimation of treatment effect at the endpoint, and a positive bias indicates an overestimation of treatment effect at the endpoint. RMSE is reported as a measure of overall accuracy of the estimate of treatment effect; this measure combines the bias and variance of the estimate. The RMSE for the data without missing values ranged from 2.67–2.81 across the six variance-covariance matrices irrespective of ANCOVA or MMRM. In the following four subsections, the results are presented under each of the four missing data mechanisms.

5.2.1 Bias and RMSE under MCAR

Figure 5.1 displays the bias in estimates of treatment effect for each of the missing data handling methods under equal and unequal dropout rate between study groups. CCA, MMRM and MI gave unbiased estimates of treatment effect irrespective of 10% or 30% dropout rate, equal or differential dropout rate between groups, level of data variability and level of correlation between repeated assessments. However, LOCF produced biased estimates of treatment effect even with 10% dropouts; the bias ranged from −2.4 to 0.6. Further, the level of bias increased with an increase in the level of dropout; with 30% dropout rate, the bias ranged from −4.5 to 0.6. Since the improvement in baseline score (i.e. reduction) was higher in the experimental group, the assumption of no change after dropout has a significant effect on the estimate in the experimental group compared to the control group. Therefore, as

expected, LOCF led to underestimation of the treatment effect under the scenarios of equal or higher dropout rate in the experimental group and overestimation under the scenarios of higher dropout rate in the control group.

Figure 5.2 displays the RMSE of the estimates of treatment effect for each of the missing data handling methods under equal and unequal dropout rate between groups. CCA, MMRM and MI exhibited similar RMSE in this simulation study irrespective of level of dropout rate, equal or differential dropout rate between groups, level of data variability and level of correlation between repeated assessments. However, in all these methods, RMSE increased with an increase in the level of overall dropout rate. RMSEs under LOCF were not consistent across the considered scenarios. LOCF overestimated the RMSE in respect of higher dropout rate in the experimental group, and the deviation from its true value was substantial at 30% dropout rate. Further, LOCF underestimated the RMSE in respect of higher dropout rate in the control group, and the deviation from its true value increased with higher level of overall dropout rate. When the dropout rate was 10% and equal between groups, RMSE was similar to the corresponding value in data without missing values; however, with 30% dropout rate, it was slightly overestimated.

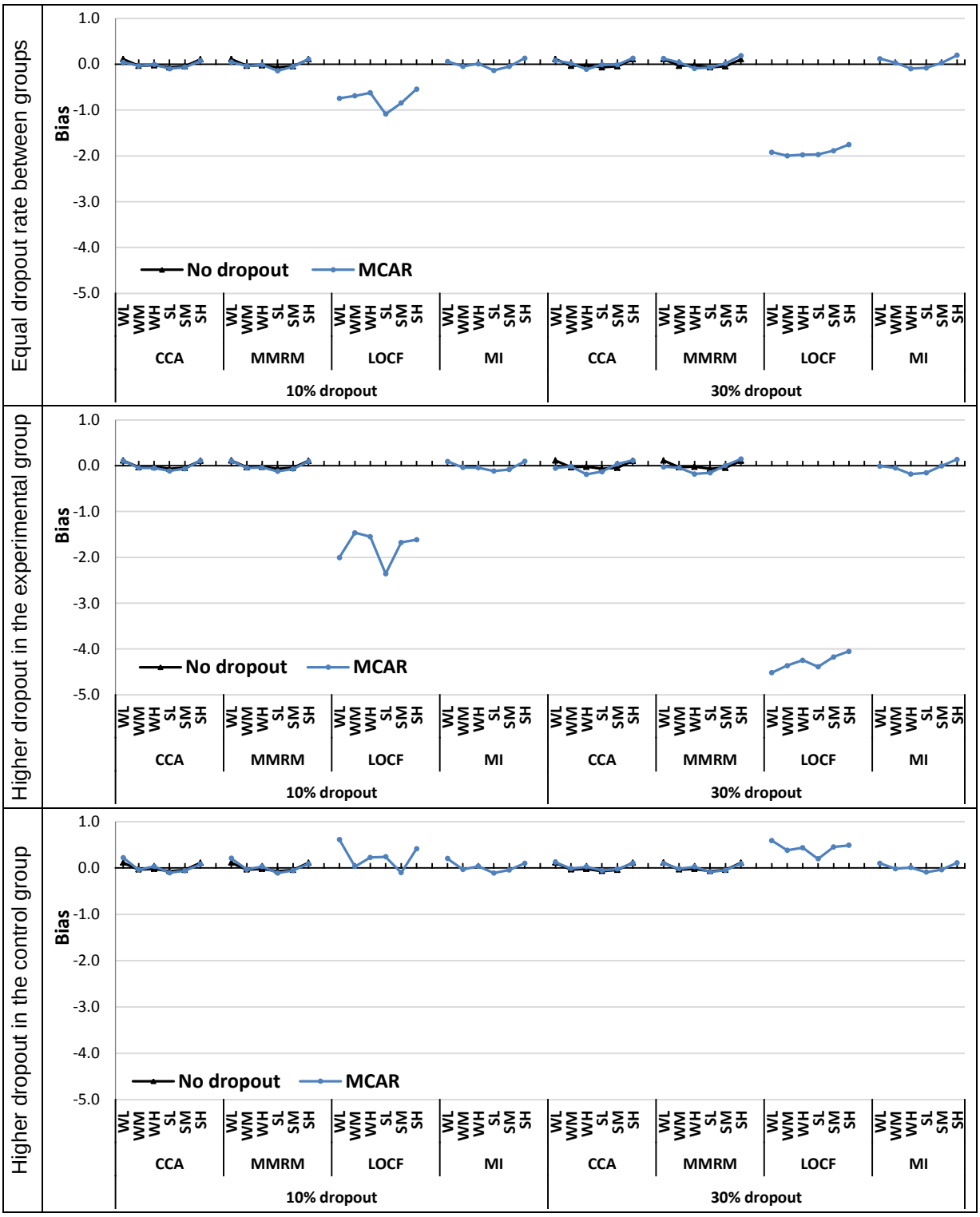


Figure 5.1: Bias under MCAR

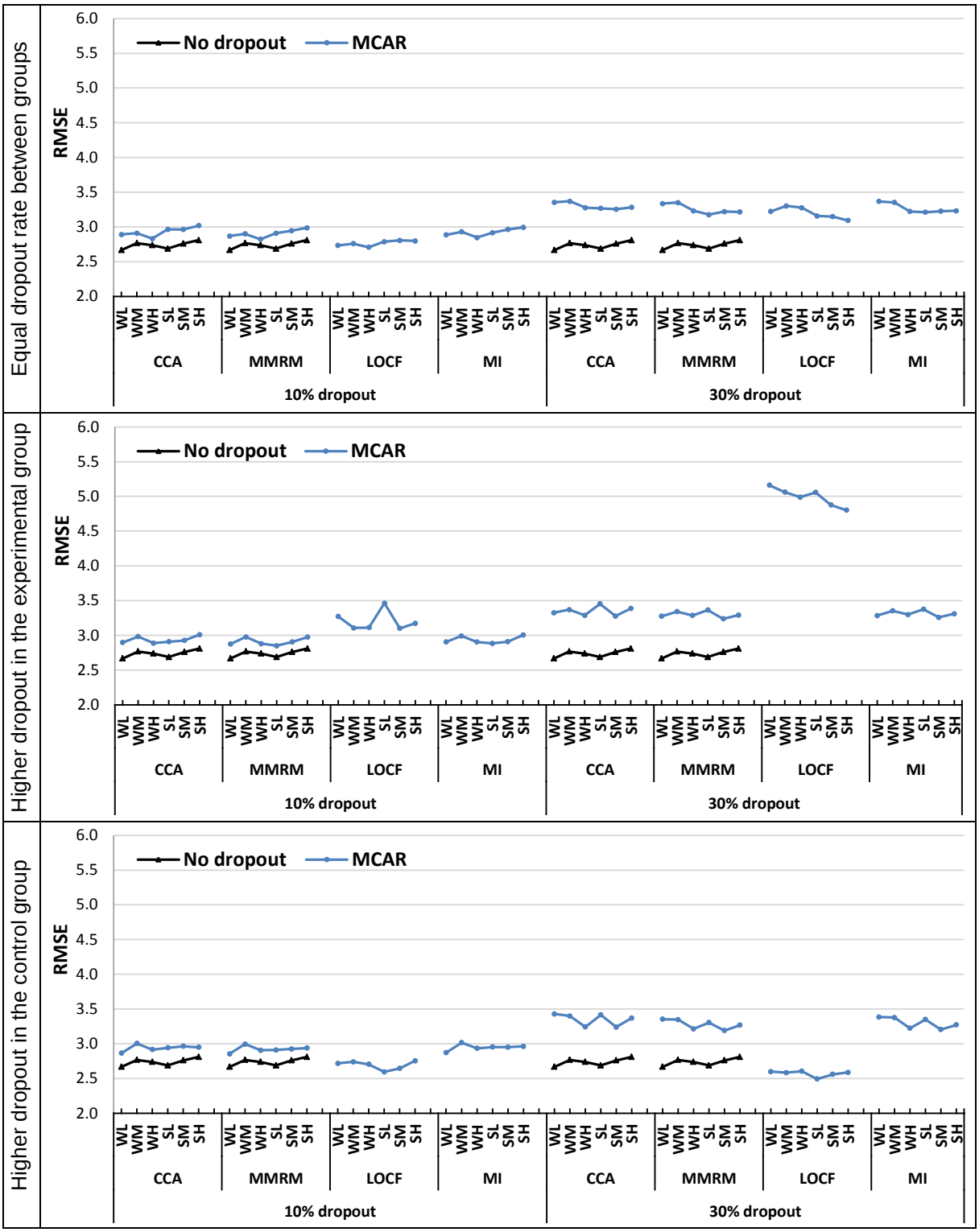


Figure 5.2: RMSE under MCAR

5.2.2 Bias and RMSE under MAR dependent on baseline value (MAR-B)

As noted before, the MAR-B mechanism was simulated under two very contrasting situations: (i) dropouts were in the same direction in both groups – participants with a high baseline score in each group were randomly dropped out (denoted as MAR-B1); (ii) dropouts were in opposite directions between the groups – participants with a high baseline score in the control group and those with a low score in the experimental group were randomly dropped out (denoted as MAR-B2). This simulation study found that the accuracy (in terms of bias and RMSE) of all the methods under MAR-B1 was similar to the situation under MCAR.

Figure 5.3 displays the bias in estimates of treatment effect for each of the missing data handling methods under equal and unequal dropout rate between groups. All approaches except LOCF were accurate in estimating the true treatment effect irrespective of level of dropout rate, equal or unequal dropout rate between groups, level of data variability, and level of correlation between repeated assessments in addition to, importantly, the direction of dropouts – whether it was in the same or opposite direction. With LOCF, the bias in estimates of treatment effect at the endpoint was substantial in most scenarios. However, the bias was not consistent across the scenarios; the absolute degree of bias (whether underestimation or overestimation) was dependent on the imbalance in dropout rate between groups. Further, the trend was stronger when the dropouts were in opposite directions between the groups and the data variability was high. The bias ranged from −2.4 to 1.8 with 10% dropout rate and from −5.1 to 3.4 with 30% dropout rate.

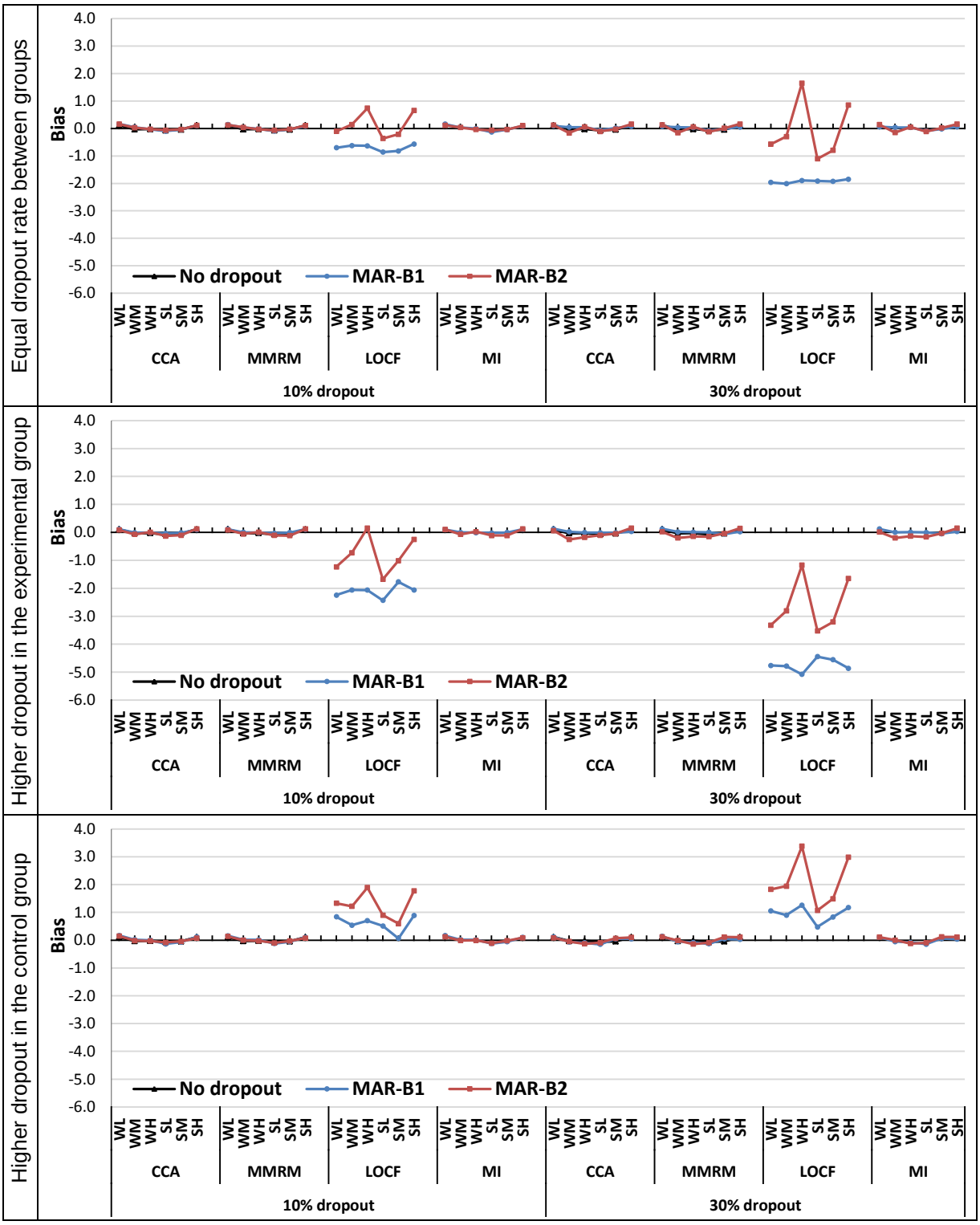


Figure 5.3: Bias under MAR-B

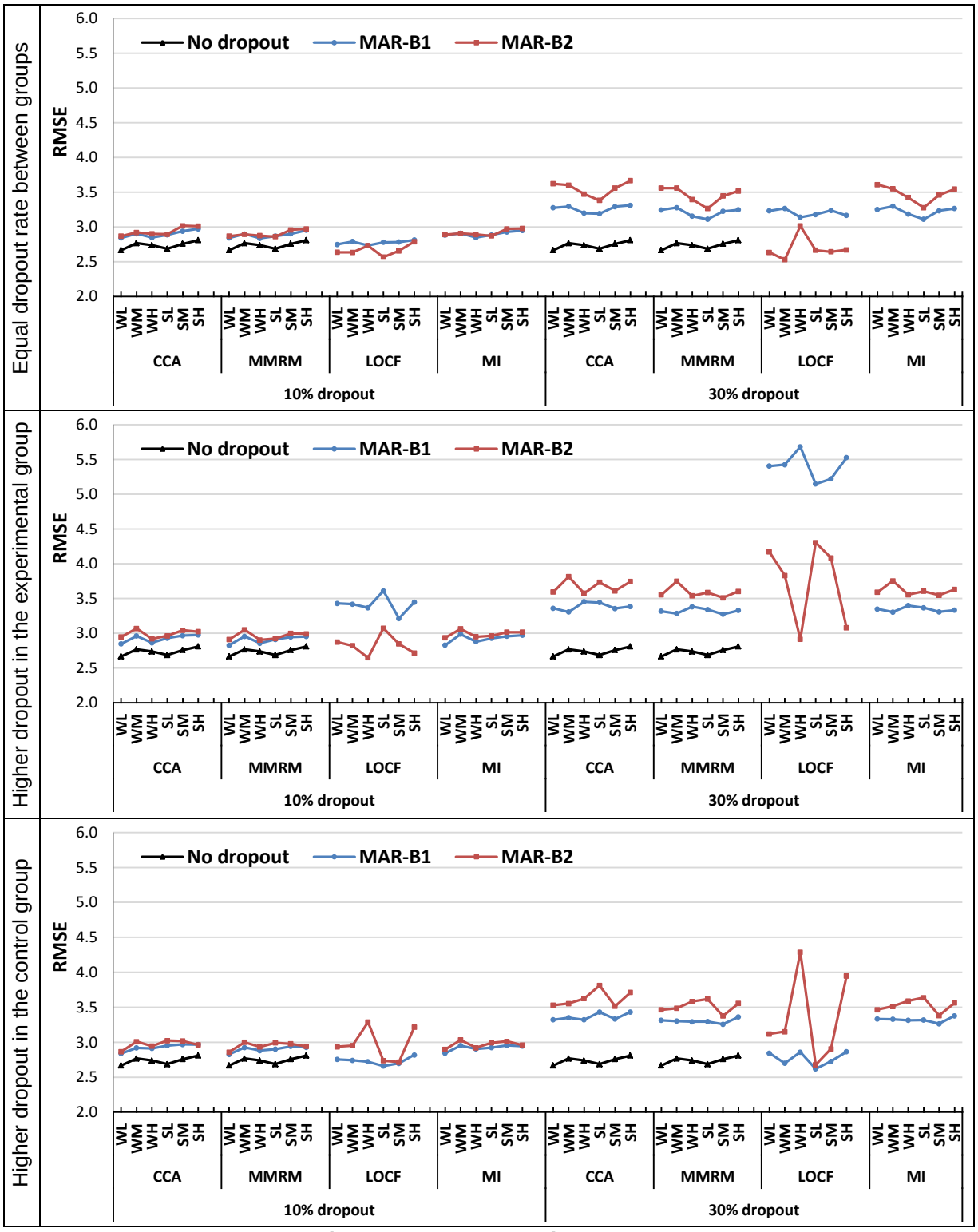


Figure 5.4: RMSE under MAR-B

Figure 5.4 displays the RMSE of the estimates of treatment effect for each of the missing data handling methods under equal and unequal dropout rate between groups. With 10% dropout rate, RMSE of the estimates under CCA, MMRM and MI were almost indistinguishable, but slightly overestimated compared to data without missing values, irrespective of equal or unequal dropout rate between groups, direction of dropouts, level of data variability, and level of correlation between repeated assessments. With 30% dropout rate, RMSE was substantially overestimated in these methods; however, it was almost identical for MMRM and MI but very slightly lower than that of CCA. Further, with 30% dropout rate, there was a noticeable difference within each of these methods in relation to direction of dropouts; higher RMSE was observed when dropouts were in opposite directions between groups. As in the case of the bias in treatment effect, RMSE under LOCF deviated considerably from the true value, especially, in relation to higher level of overall dropout, unequal dropout between groups, direction of dropouts and high level of data variability.

5.2.3 Bias and RMSE under MAR dependent on last observed value (MAR-L)

As in the case of MAR-B, the MAR-L mechanism was simulated under two very contrasting situations: (i) dropouts were in the same direction in both groups – participants with a high last observed score in each group were randomly dropped out (denoted as MAR-L1); (ii) dropouts were in opposite directions between the groups – participants with a high last observed score in the control group and those with a low last observed score in the experimental group were randomly dropped out (denoted as MAR-L2). This simulation study found that the accuracy (in terms of bias and RMSE) of all the methods under MAR-L1 was similar to the situation under MCAR when dropout was equal between groups. Further, accuracy of MMRM and MI under MAR-L1 and MCAR was similar even with differential dropout rate between groups.

Figure 5.5 displays the bias in estimates of treatment effect for each of the missing data handling methods under equal and unequal dropout rate between groups. Under situations where dropouts were in the same direction in both groups (i.e. MAR-L1) and equal dropout rate between the groups, all but LOCF approaches yielded an unbiased estimate of treatment effect irrespective of level of dropout rate, level of data variability and level of correlation between repeated assessments. However, with an unequal dropout rate between groups in this situation, CCA led to biased estimates – underestimated with higher dropout in the control group and overestimated with higher dropout in the experimental group. The bias ranged from −0.7 to 0.8 with 10% dropout rate and −2.3 to 2.4 with 30% dropout rate. When dropouts were in opposite directions between the groups (i.e. MAR-L2), CCA underestimated the treatment effect irrespective of equal or unequal dropout rate between the groups. The bias ranged from −1.6 to −0.6 with 10% dropout rate and −4.5 to −1.5 with 30% dropout rate. Further, the amount of bias was substantial with high data variability and a high dropout rate. For example, in the situation where dropouts were in opposite directions between the groups and for equal dropout rate between the groups, CCA underestimated the treatment effect by 40% when the SD was high and the dropout rate was 30%. The bias under LOCF was worse than that under MAR-B. The bias ranged from −3.1 to 4.4 with 10% dropout rate and −7.4 to 9.1 with 30% dropout rate.

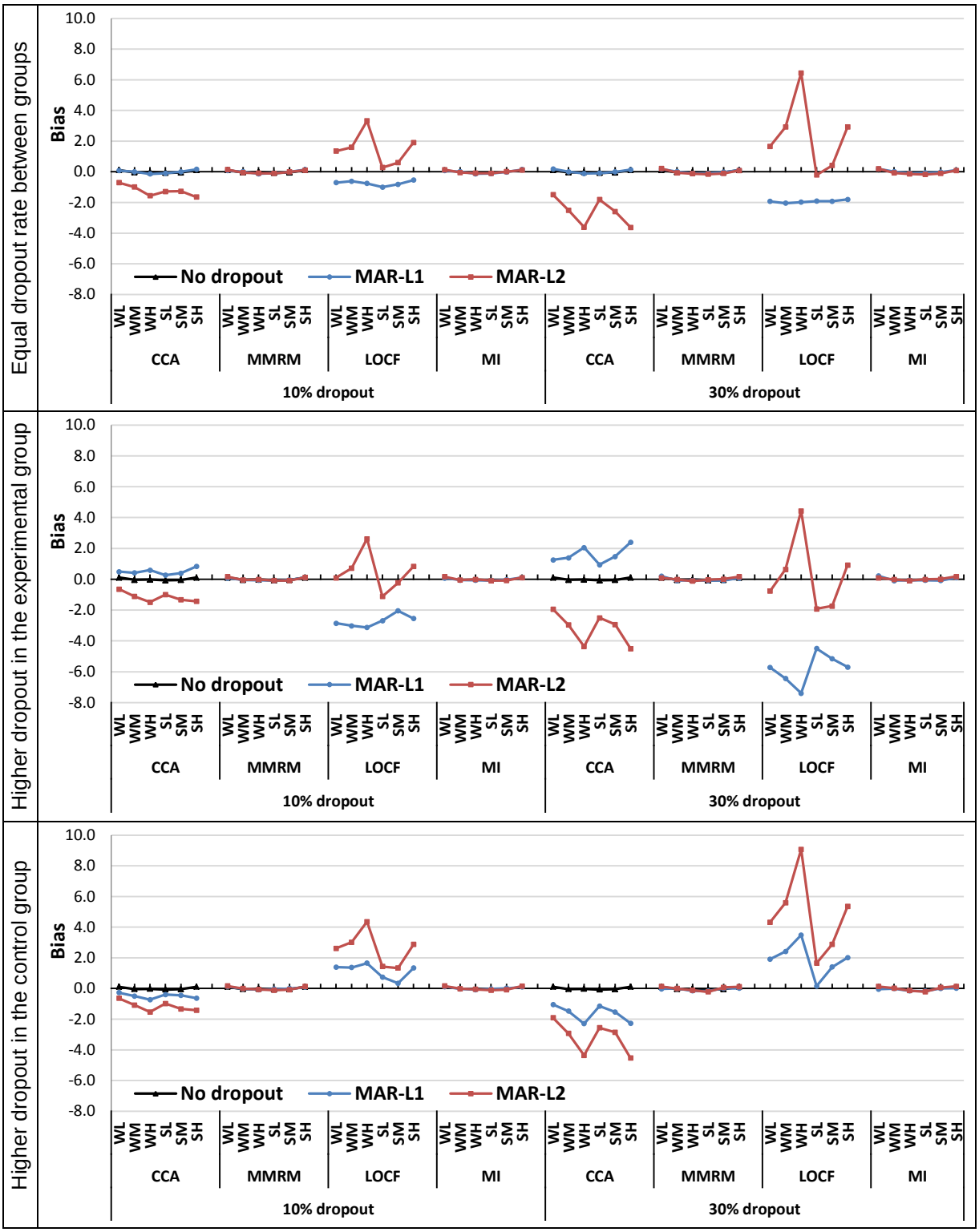


Figure 5.5: Bias under MAR-L

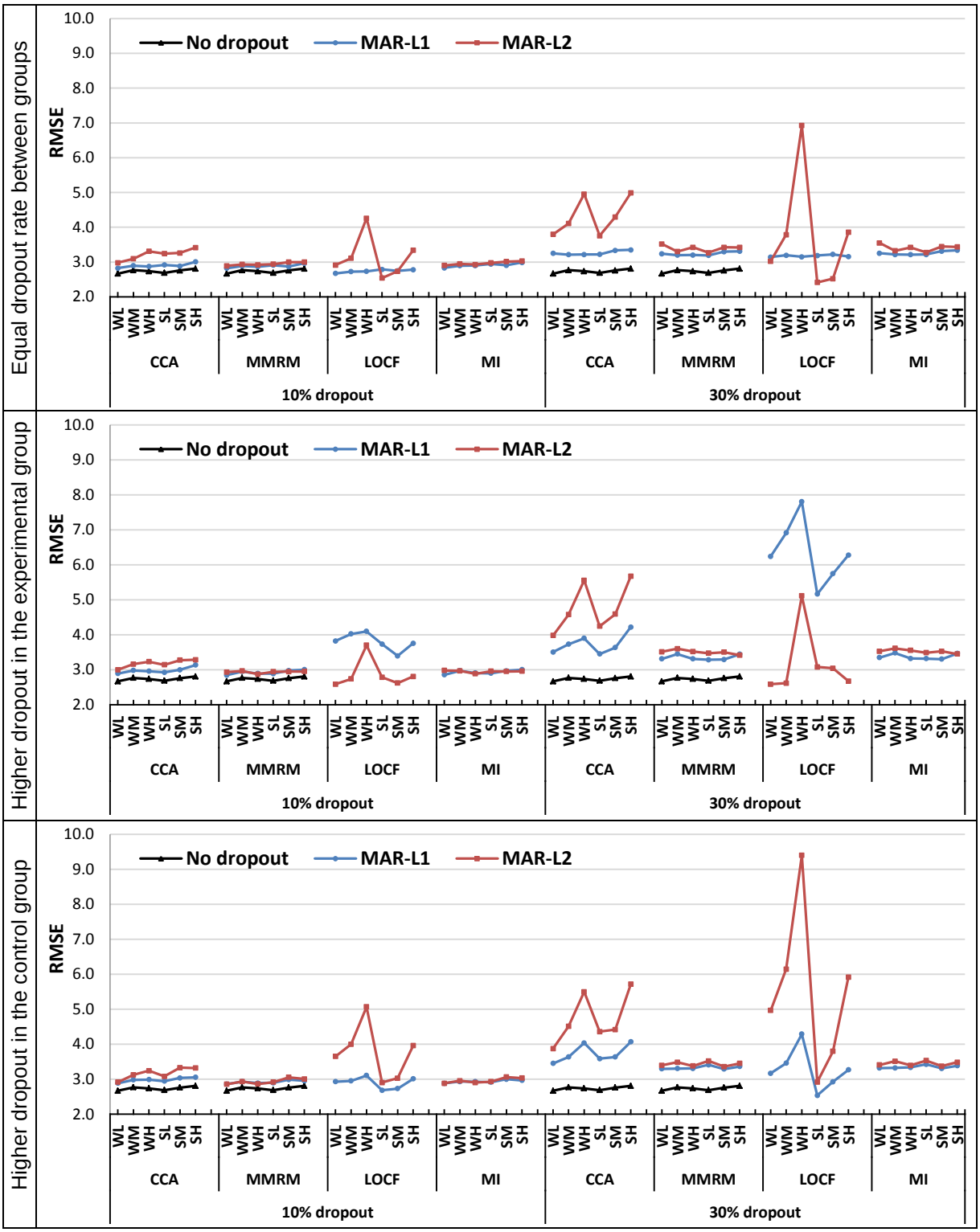


Figure 5.6: RMSE under MAR-L

Figure 5.6 displays the RMSE of the estimates of treatment effect for each of the missing data handling methods under equal and unequal dropout rate between groups. RMSE of the estimates with MMRM and MI was similar to that under MAR-B; however, very slightly lower RMSE was observed under MAR-L with 30% dropout rate. RMSE under these two methods was similar across the levels of data variability, correlation between repeated assessments and dropout rate between the groups, but slightly increased in relation to increased dropout rate. Unlike the situation in MAR-B, CCA led to a substantial increase in RMSE with 30% overall dropout rate and, especially, with dropouts in opposite directions. The inflation in RMSE was furthermore associated with greater data variability and level of correlation between the repeated assessments. As in the case of the bias in estimate of treatment effect, LOCF markedly affected the RMSE even with 10% dropout rate.

5.2.4 Bias and RMSE under MNAR

As noted before, the MNAR mechanism was simulated under two very contrasting situations: dropouts were in the same direction in both groups – participants with a high score at the time of dropping out in each group were dropped out (denoted as MNAR-1); dropouts were in opposite directions between the groups – participants with a high score at the time of dropping out in the control group and those with a low score in the experimental group were dropped (denoted as MNAR-2). This simulation study found that the accuracy (in terms of bias and RMSE) of all the methods under MNAR-1 was similar to the situation under MCAR when dropout rate was equal between groups.

Figures 5.7a and 5.7b display the bias in estimates of treatment effect for each of the missing data handling methods under equal and unequal dropout rate between groups. Figure 5.7a displays the bias in relation to level of dropout rate with a fixed strong correlation and moderate SD. It can be seen that all methods displayed an increase in magnitude of the bias

in proportion to the level of dropout rate with the exception of MI, MMRM and CCA when dropout was equal and in the same direction for the two treatment groups. Level of bias was certainly evident when the dropouts were in opposite directions between the groups. With 10% dropout rate, the bias ranged from −2.9 to 1.0 for CCA, −2.2 to 0.8 for MMRM and MI, and −2.6 to 0.4 for LOCF; with 30% dropout rate, it was −6.4 to 3.4, −4.9 to 2.6 and −6.1 to −0.4, respectively. Figure 5.7b displays the bias in relation to data variability and correlation between repeated assessments with 30% dropout rate. It can be seen that all the approaches displayed an increase in magnitude of the bias in relation to the data variability but the increase was slightly limited by a strong correlation, with the exception of MI, MMRM and CCA when dropout was equal and in the same direction for the two treatment groups. The bias became severe with an MNAR-2 mechanism. With a weak correlation and moderate SD, the bias ranged from −7.9 to 4.3 for CCA, −7.0 to 3.9 for MMRM and MI, and −7.6 to −1.3 for LOCF; with a strong correlation, it was −6.4 to 3.4, −4.9 to 2.6 and −6.1 to −0.4, respectively. These findings indicated a possible lower bias with MMRM and MI under strong correlation.

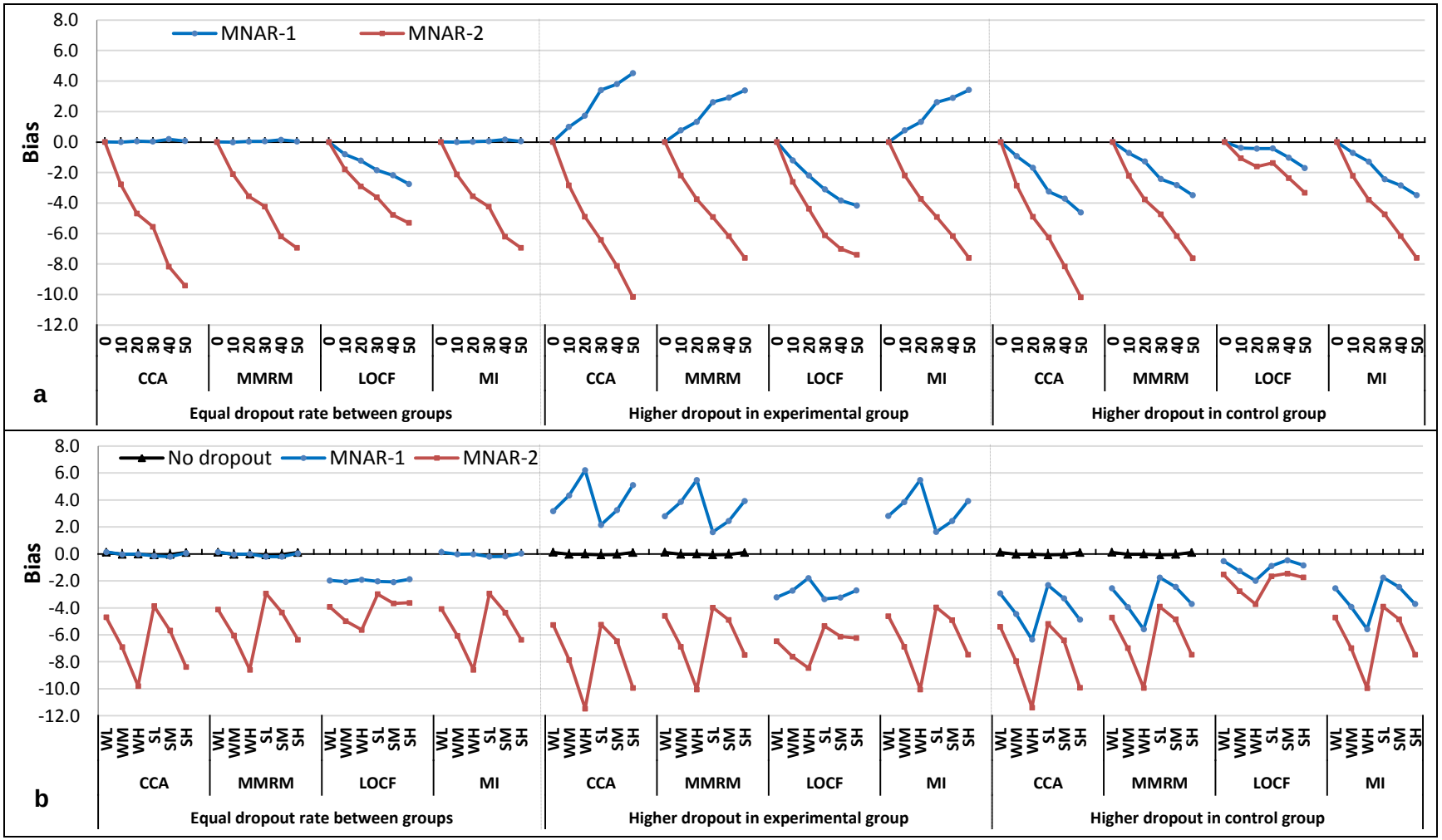


Figure 5.7: Bias under MNAR – in relation to (a) level of dropout rate (%) with a fixed strong correlation and moderate SD; (b) data variability and correlation between the repeated assessments with 30% dropout rate

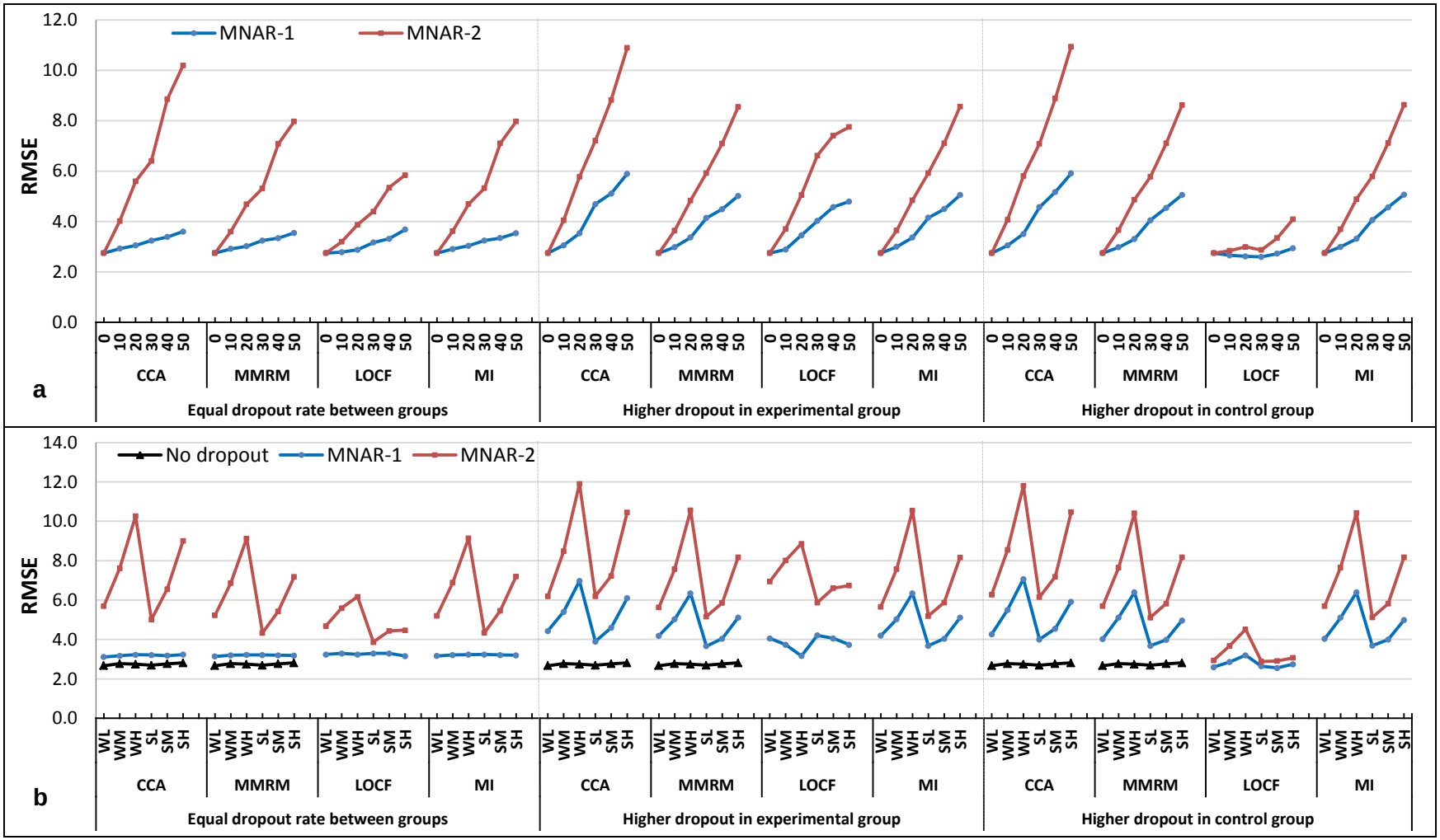


Figure 5.8: RMSE under MNAR – in relation to (a) level of dropout rate (%) with a fixed strong correlation and moderate SD; (b) data variability and correlation between the repeated assessments with 30% dropout rate

Figures 5.8a and 5.8b display the RMSE of the estimates of treatment effect for each of the missing data handling methods under equal and unequal dropout rate between groups. Figure 5.8a displays the RMSE in relation to level of dropout rate with a fixed strong correlation and moderate SD. As in the case of bias, RMSEs were also substantially inflated in proportion to the dropout rate, and the issue was more severe when the dropouts were in opposite directions between the groups. Figure 5.8b displays the RMSE in relation to data variability and correlation between the repeated assessments with 30% dropout rate. As expected, RMSEs in all methods were affected by the level of data variability but slightly controlled by a strong correlation, with the exception of MI, MMRM and CCA when dropout was equal and in the same direction for the two treatment groups. However, the inflation in RMSE under MMRM and MI was restricted by the strong correlation between repeated assessments compared to CCA.

5.3 Confidence interval coverage and width

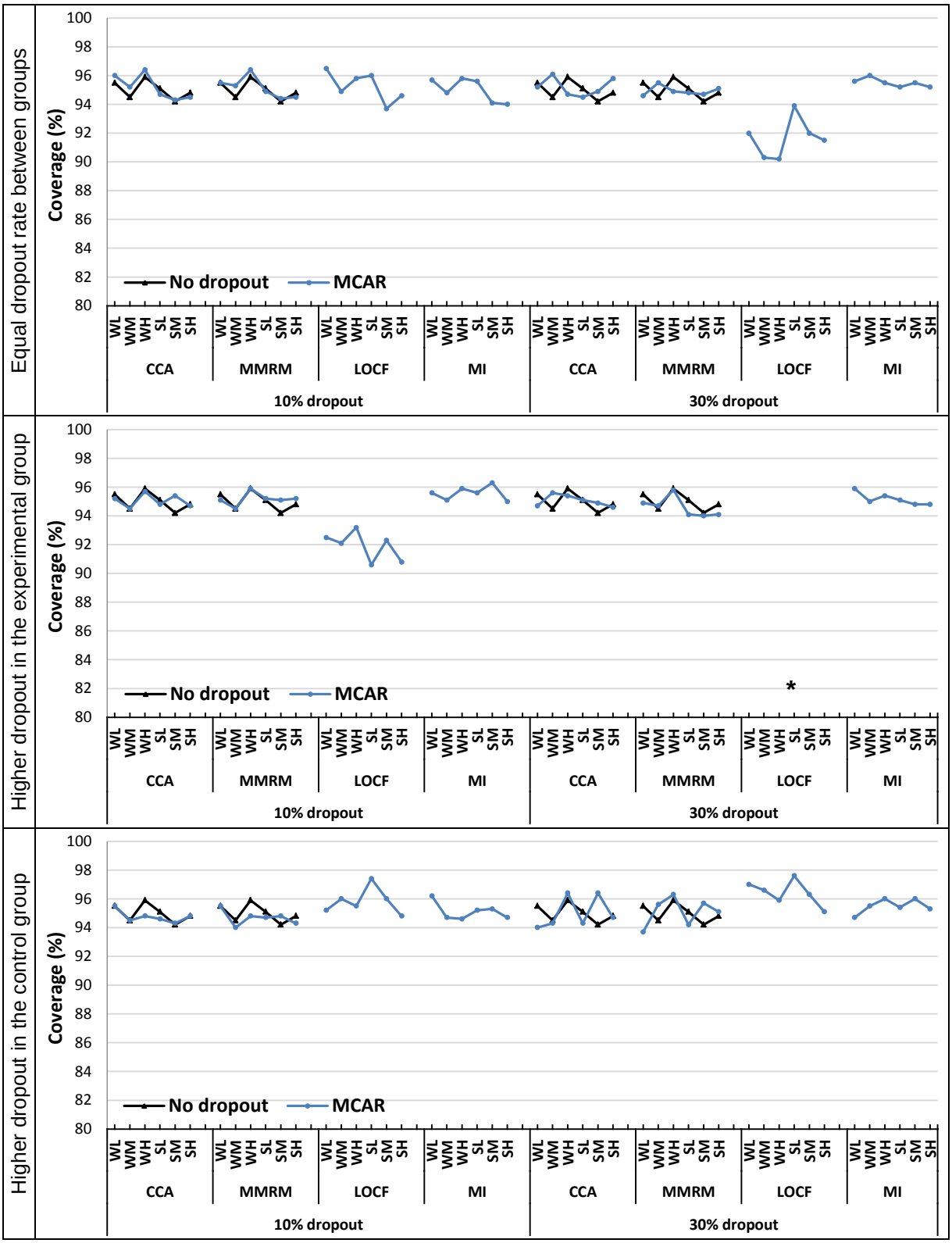
In this section, the simulation results are presented in terms of coverage and width of 95% CI under various scenarios within each missing data mechanism. As Burton et al. (2006) suggested the observed coverage that falls inside the interval 93.6%–96.4% is considered to be an acceptable coverage of the nominal 95% CI. If the coverage rate is accurate (i.e. close to 95%), the probability of a type I error will also be accurate in accordance with the designated 5% nominal level (Burton et al., 2006; Collins et al., 2001). Subject to correct coverage, confidence interval of an estimate of treatment effect should be narrow, because the smaller interval will reduce the probability of a type II error (Burton et al., 2006). The width of the CI for data without missing values ranged between 10.68–10.89 across the six variance-covariance matrices irrespective of ANCOVA or MMRM. The observed difference in the width between the variance-covariance matrices was mainly due to random sample error

since different sample sizes had been used to generate the complete data in order to ensure 90% power. In the following four subsections, the results are presented under each of the four missing data mechanisms.

5.3.1 CI coverage and width under MCAR

Figure 5.9 displays the CI coverage of each of the missing data handling methods under equal and unequal dropout rate between groups. For all approaches other than LOCF, the observed coverage of the 95% CI was within the acceptable range (i.e. 93.6%–96.4%) irrespective of all considered scenarios. However, even with 10% overall dropout rate, LOCF led to over-coverage when there was a higher dropout rate in the control group, and otherwise led to under-coverage; the situation was much worse with 30% dropout rate. The coverage of LOCF under the different scenarios considered ranged from 90.6%–97.4% with 10% dropout rate and from 66.0%–97.6% with 30% dropout rate. Further, LOCF showed a trend of reduction in coverage in relation to increase in data variability and correlation between repeated assessments.

Figure 5.10 displays the average CI width in each of the missing data handling methods under equal and unequal dropout rate between groups. As noted earlier, for data without missing values, the width of the 95% CI was approximately 10.8 units irrespective of the data characteristics (data variability and correlation between repeated assessments). LOCF was the approach with the smallest CI width among the four approaches; however, with a high SD the width was narrower than that for the data without missing values – indicating an underestimation of the standard error. The width varied by the level of data variability but was fairly consistent across different dropout rates, being in the range of 10.8 to 11.7 with 10% dropout rate and 10.5 to 11.9 with 30% dropout rate. In all the other approaches, the width increased in proportion to the level of dropout rate.



*coverage was less than 80%

Figure 5.9: CI coverage under MCAR

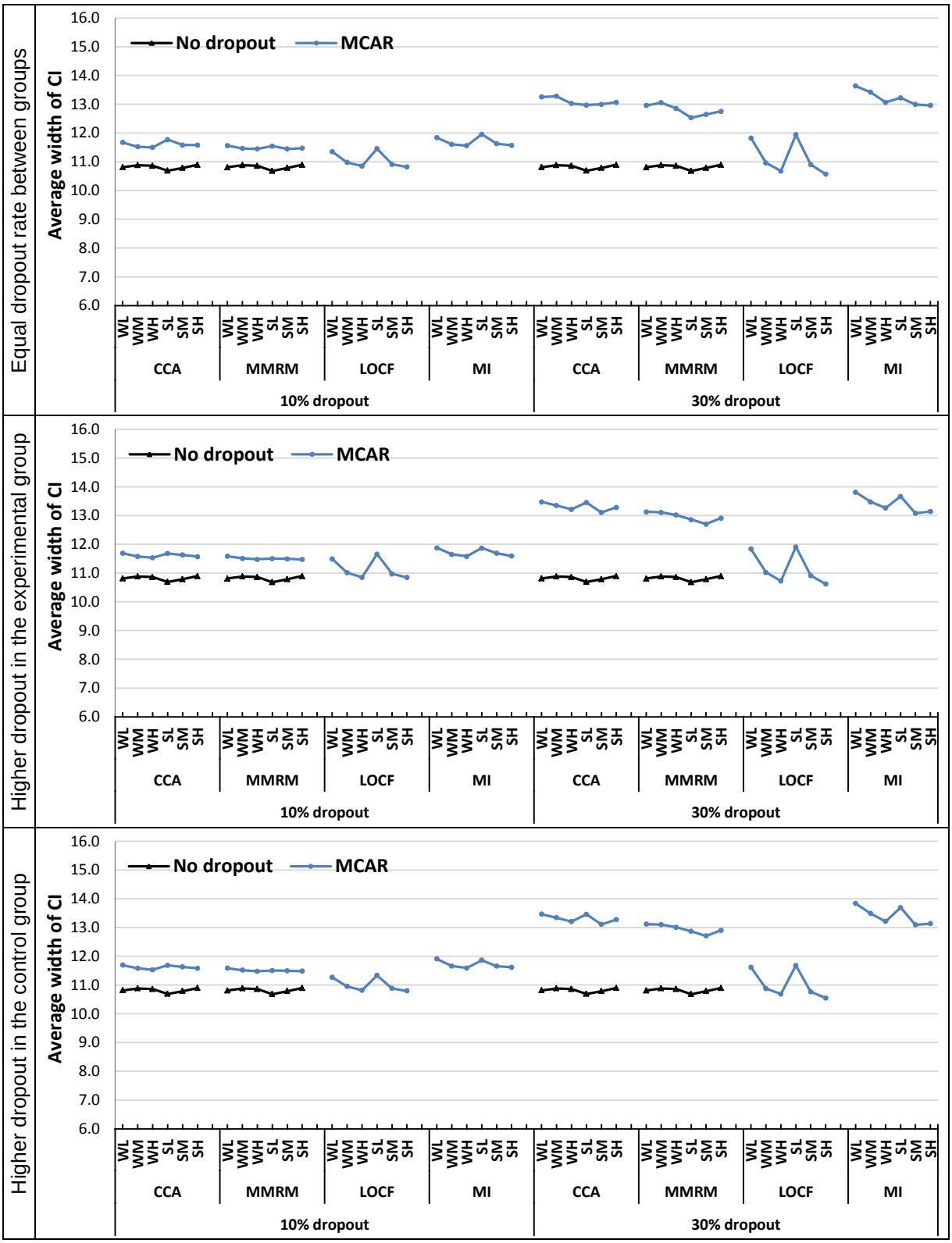


Figure 5.10: Average width of the 95% CI under MCAR

The width in CCA and MMRM was unaffected by data variability, correlation and dropout rate between the groups; the width in MI was slightly affected by the data variability and the correlation (i.e. the width was slightly narrower with a high SD and a strong correlation). With 10% dropout rate, the width ranged from 11.5–11.8 for CCA, from 11.5–11.6 for MMRM and from 11.6–12.0 for MI; with 30% dropout rate, it was 13.0–13.5, 12.5–13.1 and 13.0–13.8, respectively. In general, the CI width was slightly narrower with MMRM compared to CCA and MI.

5.3.2 CI coverage and width under MAR-B

Figure 5.11 displays the CI coverage of each of the missing data handling methods under equal and unequal dropout rate between groups. As in the case of MCAR, all but the LOCF approach retained the coverage within the acceptable range nearly in all scenarios irrespective of level of overall dropout rate, dropout rate between groups, direction of dropouts, data variability and correlation between repeated assessments. With 10% dropout rate, the observed coverage ranged from 94.0%–96.9% for CCA, from 93.7%–96.4% for MMRM and from 94.3%–96.3% for MI; with 30% dropout rate, it was from 93.3%–96.6%, 93.2%–96.0% and 93.8%–96.6% respectively. However, deviation from the nominal coverage was of real concern in LOCF, even with a 10% dropout rate; in most scenarios, LOCF led to under-coverage, and in some scenarios it was worse than 60%. The deviation was substantially dependent on the data variability, level of overall dropout, dropout rate between groups and direction of dropouts. The observed coverage with LOCF ranged from 88.7%–97.2% with 10% dropout rate and from 55.8%–98.0% with 30% dropout rate.

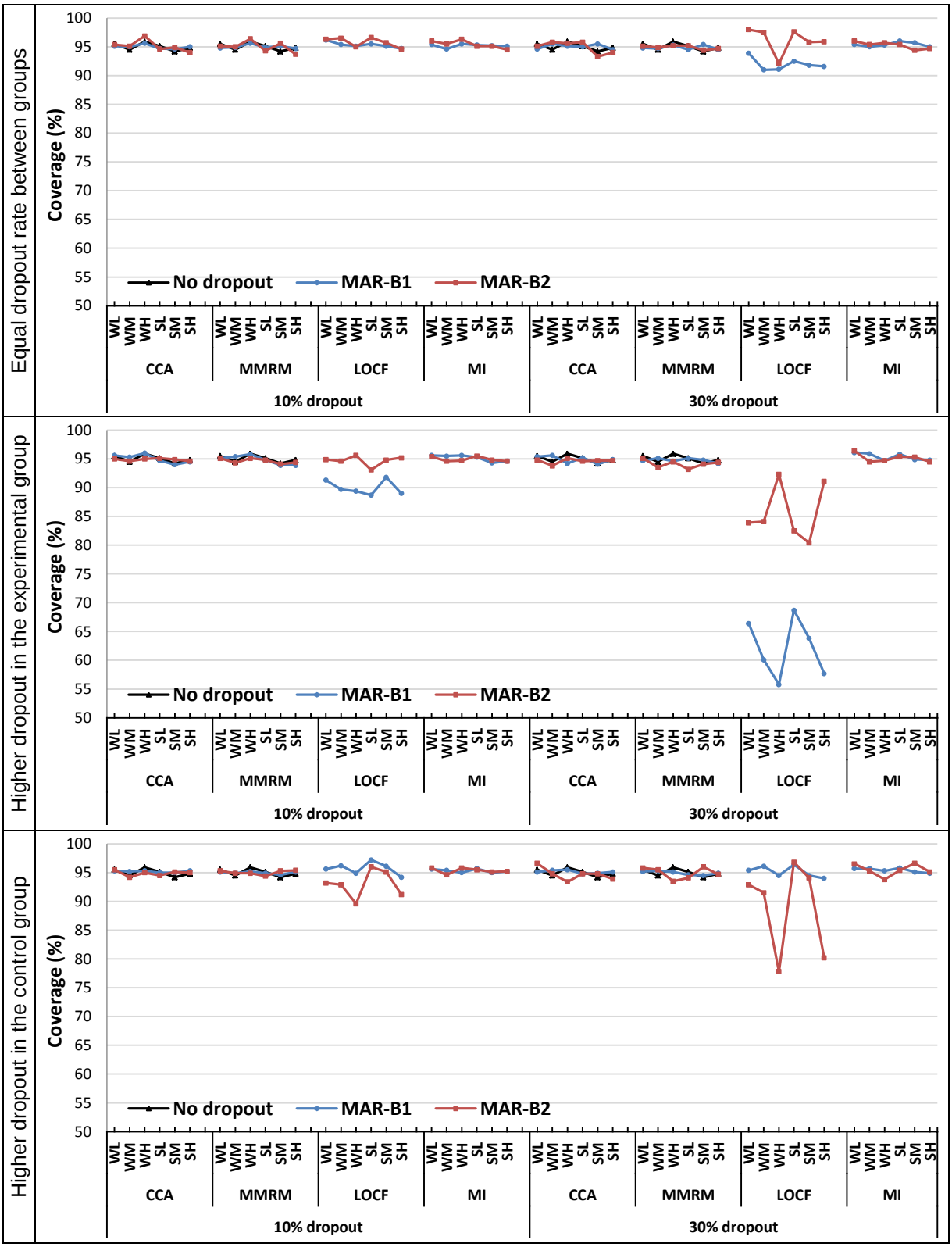


Figure 5.11: CI coverage under MAR-B

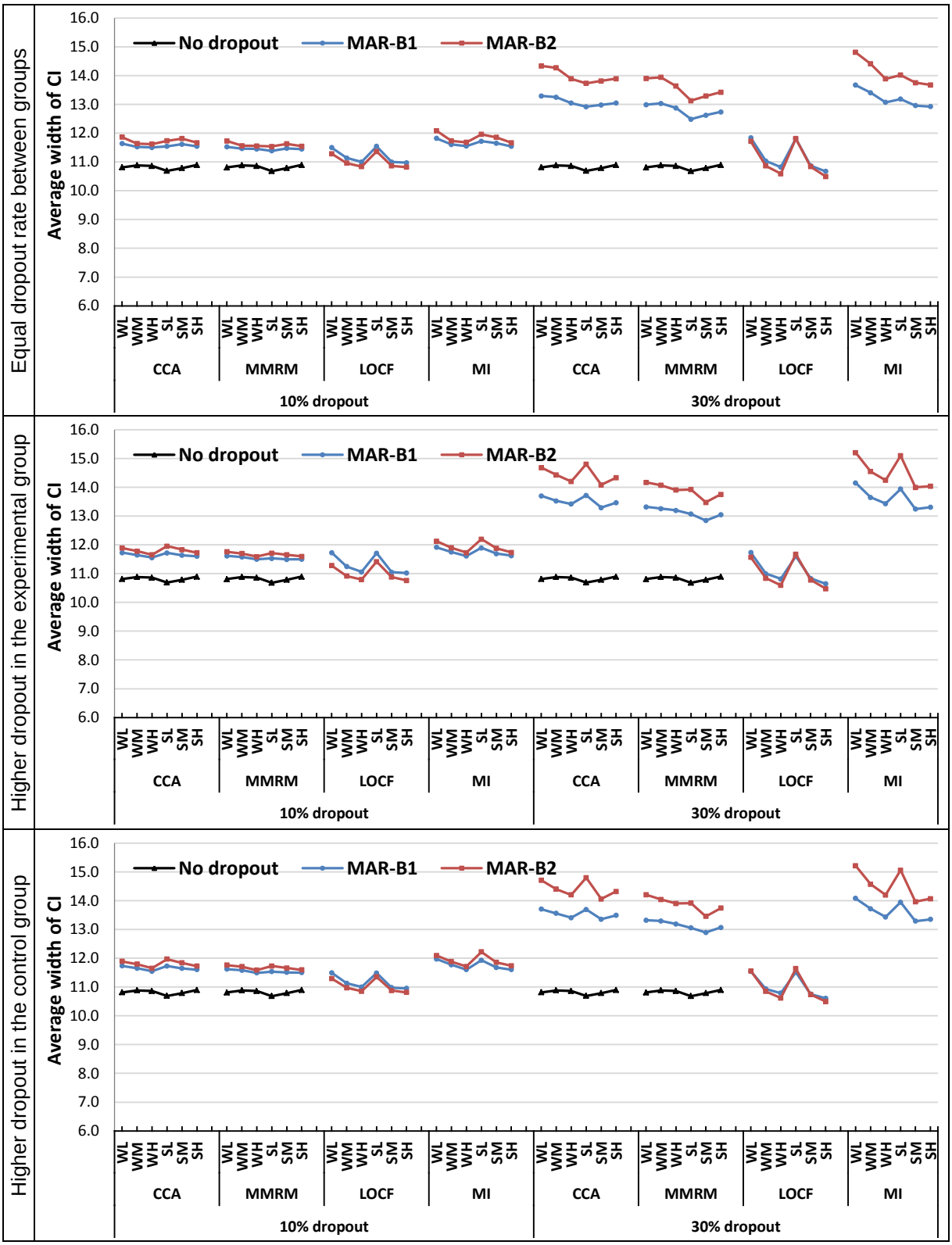


Figure 5.12: Average width of the 95% CI under MAR-B

Figure 5.12 displays the average CI width in each of the missing data handling methods under equal and unequal dropout rate between groups. When the dropouts were in the same direction in both groups (i.e. MAR-B1), the findings were similar to the situation under MCAR. From the figure, it can be seen that LOCF was the approach with the narrowest CI width; however, it led to a narrower CI compared to data without missing values when the data variability was high – indicating an underestimation of the standard error. Though the width under LOCF was associated with data variability, it was less affected by the level of dropout rate, equal or unequal dropout rate between groups, direction of dropout and correlation between repeated assessments. The width ranged from 10.8–11.7 with 10% dropout rate and from 10.5–11.8 with 30% dropout rate – which was similar to the situation under MCAR. In all the other approaches, the width increased in proportion to the overall dropout rate, and was further influenced by direction of dropout rate and data variability when the dropout rate was high. It was observed that the data variability had slightly more impact on the width in MI compared to MMRM. In addition, MMRM produced a slightly narrower CI compared to CCA and MI. With 10% dropout rate, the width ranged from 11.5–12.0 for CCA, from 11.4–11.8 for MMRM and from 11.5–12.2 for MI; with 30% dropout rate, it was 12.9–14.8, 12.5–14.2 and 13.0–15.2, respectively.

5.3.3 CI coverage and width under MAR-L

Figure 5.13 displays the CI coverage of each of the missing data handling methods under equal and unequal dropout rate between groups. Under the situation where dropouts were in the same direction in both groups (i.e. MAR-L1) and the dropout rate was equal between the groups, all but the LOCF approaches yielded similar CI coverage – and the observed coverage was within the acceptable range – irrespective of level of dropout rate, level of data variability and level of correlation between repeated assessments.

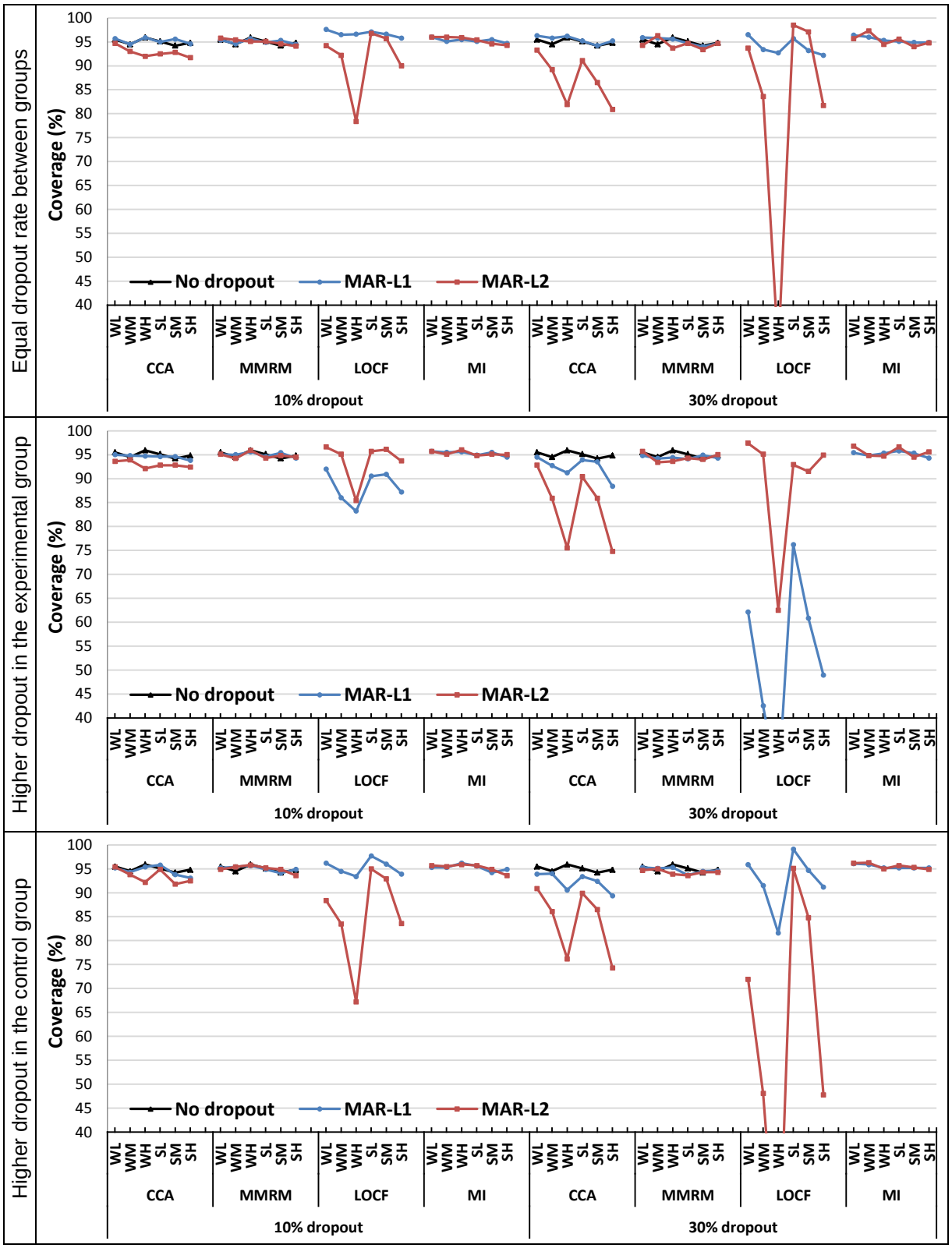


Figure 5.13: CI coverage under MAR-L

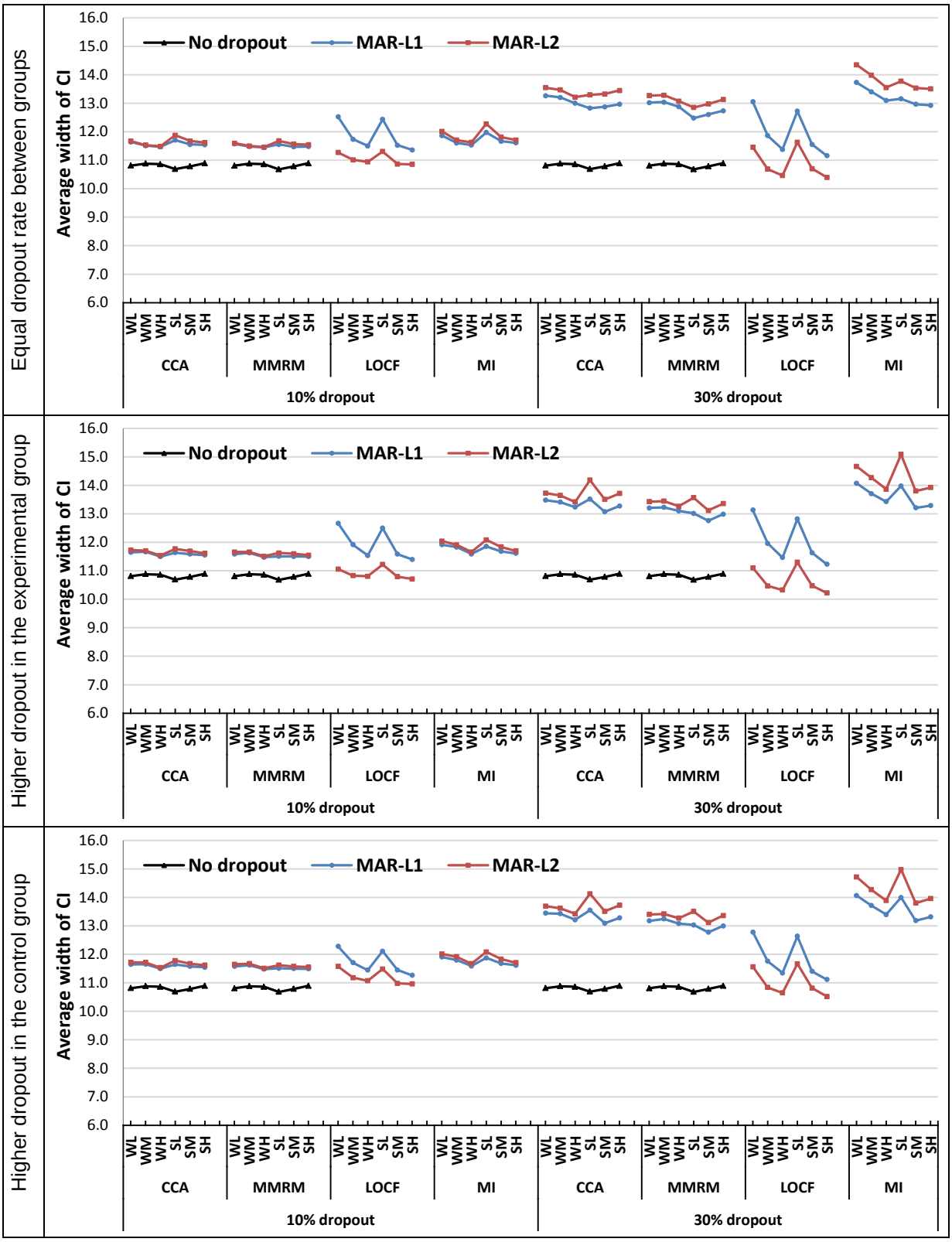


Figure 5.14: Average width of the 95% CI under MAR-L

However, with an unequal dropout rate between groups, CCA led to under-coverage; the loss of coverage was noticeable with 30% dropout rate. When dropouts were in opposite directions between the groups (i.e. MAR-L2), CCA led to under-coverage irrespective of equal or unequal dropout rate between the groups; the loss of coverage was substantial with 30% dropout rate and further influenced by the data variability. The observed coverage with MMRM and MI was within the acceptable range in most scenarios; however, as in MAR-B, MI retained very slightly higher coverage than that for MMRM. With 10% dropout rate, the observed coverage ranged from 91.7%–95.9% for CCA, from 93.6%–95.9% for MMRM and from 93.6%–96.2% for MI; with 30% dropout rate, it was 74.3%–96.3%, 93.4%–96.3% and 94.0%–97.3%, respectively. The observed coverage under LOCF severely deviated from the acceptable range, and was worse than under MAR-B. For different scenarios, it ranged from 67.2%–97.7% with 10% dropout rate and from 6.9%–99.1% with 30% dropout rate.

Figure 5.14 displays the average CI width in each of the missing data handling methods under equal and unequal dropout rate between groups. When dropouts were in the same direction in both groups (i.e. MAR-L1), the findings were similar to the situation under MCAR. Unlike the findings under MAR-B, LOCF produced a smaller CI when the dropouts were in opposite directions between groups (i.e. MAR-L2) compared to the situation where dropouts were in the same direction in both groups; the width was further influenced by the data variability. However, the observed width from LOCF was unrelated to the situation where equal or unequal dropout rate between the groups, and was less influenced by the overall dropout rate compared to other approaches. The width ranged from 10.7–12.7 with 10% dropout and from 10.2–13.1 with 30% dropout. For all other approaches, the average 95% CI became wider when the overall dropout rate was increased from 10% to 30%. The direction of dropouts also had an impact on the width – CCA, MMRM and MI produced a wider CI when the dropouts were in opposite directions between groups; however, the impact was

lower compared to the situation under MAR-B. In addition, the width with MMRM was slightly smaller than for CCA and MI. Further, the width under MI altered notably with greater data variability. With 10% dropout rate, the width ranged from 11.5–11.9 for CCA, from 11.5–11.7 for MMRM and from 11.5–12.3 for MI; with 30% dropout rate, it was 12.9–14.2, 12.5–13.6 and 12.9–15.1, respectively.

5.3.4 CI coverage and width under MNAR

Figures 5.15a and 5.15b display the observed CI coverage of each of the missing data handling methods under equal and unequal dropout rate between groups. Figure 5.15a displays the coverage in relation to level of dropout rate with a fixed strong correlation and moderate SD. It can be seen that all the approaches displayed substantial loss of coverage in proportion to increased dropout rate; the problem of loss of coverage was severe when the dropouts were in opposite directions between the groups, and this was true even with a 10% dropout rate. With 10% dropout, the coverage ranged from 83.5%–95.0% for CCA, from 87.6%–95.3% for MMRM, from 84.1%–97.0% for LOCF, and from 87.9%–94.8% for MI; with 30% dropout rate, it was 50.7%–94.7%, 66.9%–94.0%, 35.8%–97.0% and 68.6%–94.5%, respectively. Figure 5.15b displays the coverage in relation to data variability and correlation between the repeated assessments with 30% dropout. It can be seen that all the approaches displayed increasing loss of coverage in relation to greater data variability and weaker correlation, with the exception to the situation where dropout was equal and in the same direction for the groups. With a weak correlation and moderate SD, the coverage ranged from 33.5%–95.7% for CCA, from 41.1%–95.6% for MMRM, from 16.2%–95.0% for LOCF, and from 44.3%–96.0% for MI. In these two figures, it can be seen that the loss of coverage was comparatively lower in all approaches when the dropouts were in the same direction in comparison with dropouts in opposite directions between the groups.

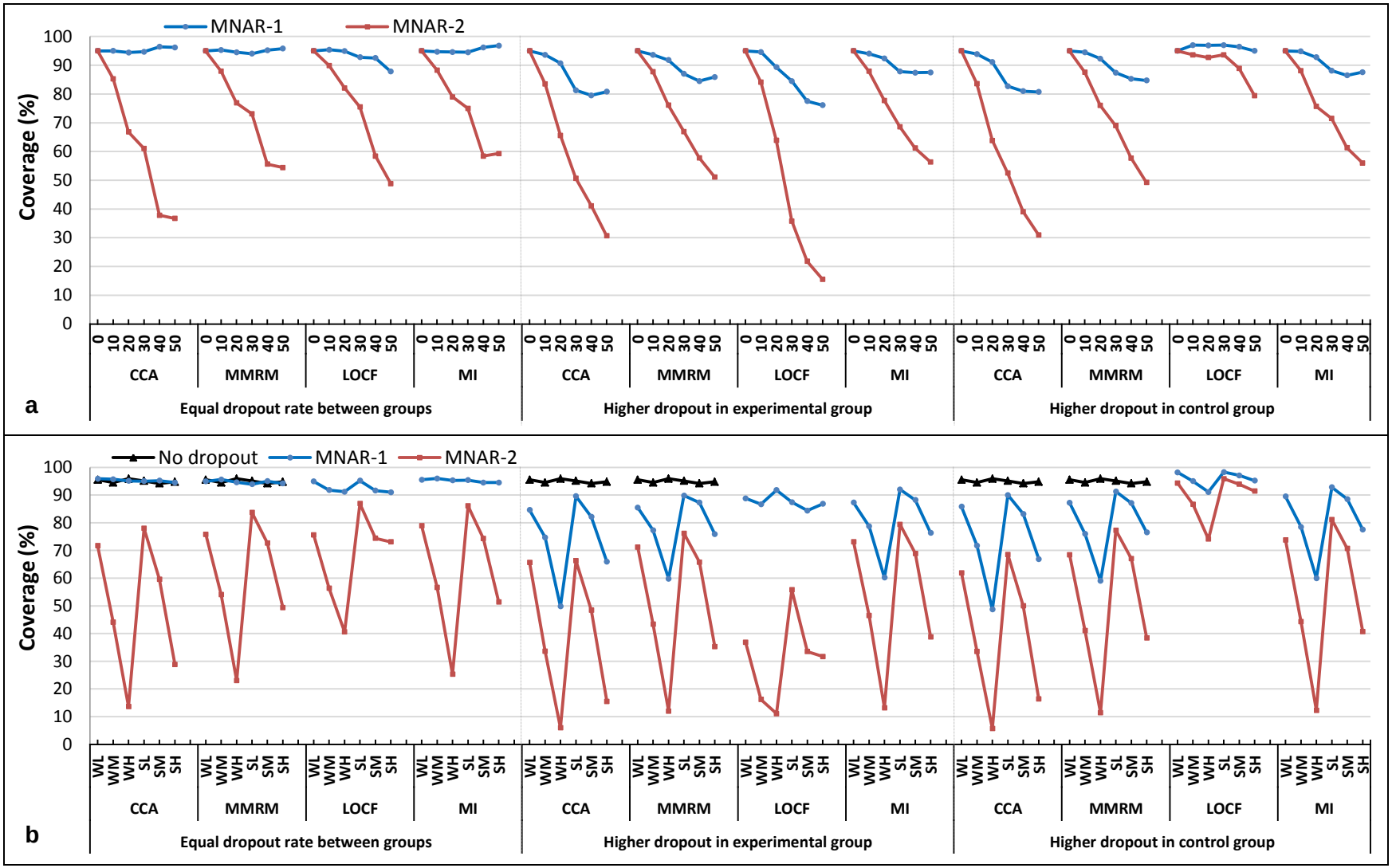


Figure 5.15: CI coverage under MNAR – in relation to (a) level of dropout rate (%) with a fixed strong correlation and moderate SD; (b) data variability and correlation between the repeated assessments with 30% dropout rate

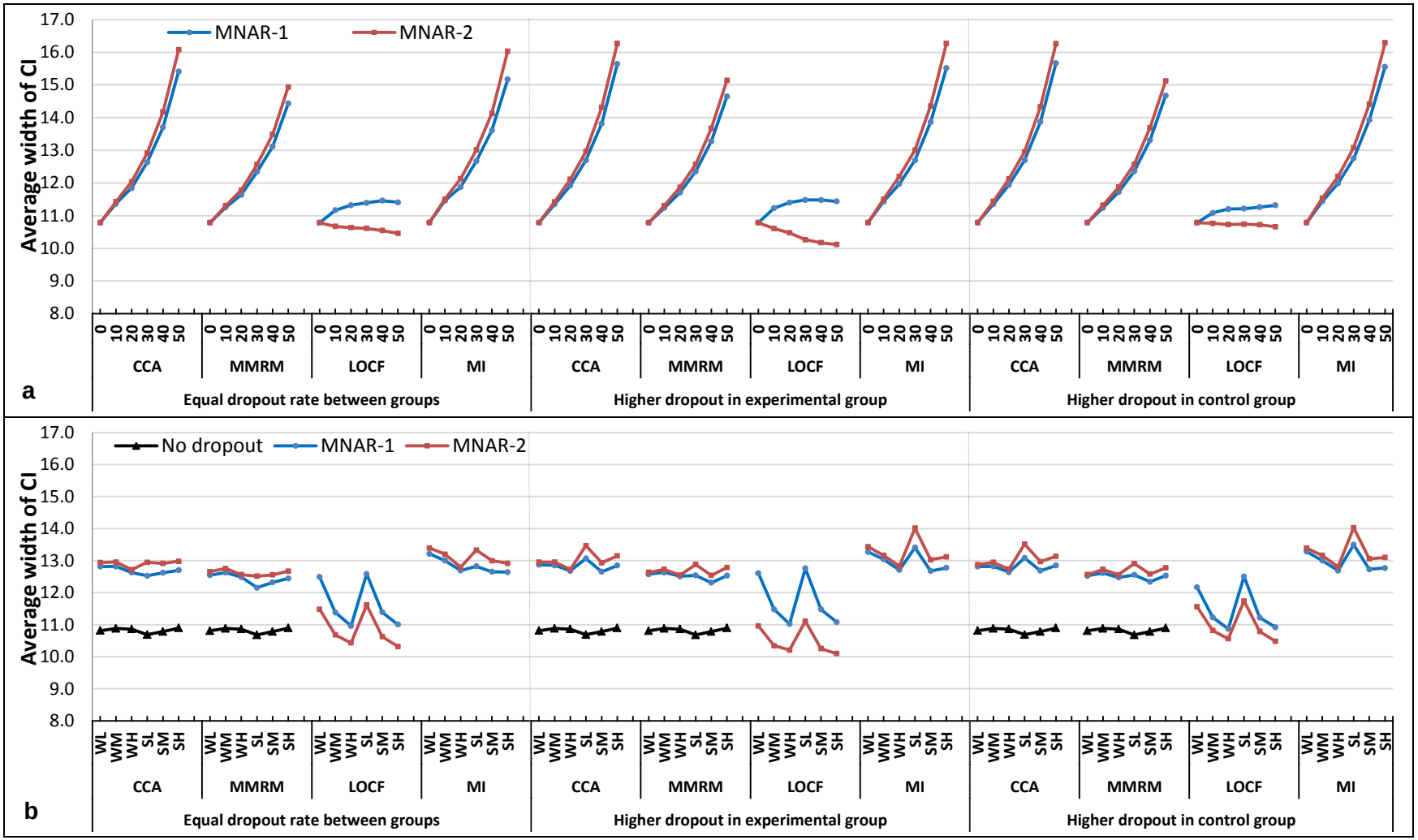


Figure 5.16: Average width of CI under MNAR – in relation to (a) level of dropout rate (%) with a fixed strong correlation and moderate SD; (b) data variability and correlation between the repeated assessments with 30% dropout rate

Figures 5.16a and 5.16b display the average CI width in each of the missing data handling methods under equal and unequal dropout rate between groups. Figure 5.16a displays the width in relation to level of dropout rate with a fixed strong correlation and moderate SD. For all but the LOCF approach, the 95% CI became wider in relation to increase in the overall dropout rate; however, the width was unaffected by the direction of dropouts and the dropout rate between the groups. LOCF produced the narrowest CI; however, it led to underestimation of standard error for many scenarios especially under MNAR-2. With 10% dropout rate, the width ranged from 11.3–11.4 for CCA, from 11.2–11.3 for MMRM, from 10.6–11.2 for LOCF, and from 11.4–11.5 for MI; with 30% dropout rate, it was 12.6–13.0, 12.3–12.6, 10.3–11.5 and 12.7–13.1, respectively. Figure 5.16b displays the width in relation to data variability and correlation between repeated assessments with 30% dropout rate. Under LOCF with 30% dropout rate, the width was markedly reduced in relation to increase in data variability. With a weak correlation and moderate SD, the coverage ranged from 12.8–13.0 for CCA, from 12.6–12.8 for MMRM, from 10.3–11.5 for LOCF, and from 13.0–13.2 for MI – indicating a slightly narrower CI under MMRM compared to MI.

5.4 Statistical power to detect the true difference

In this section, the simulation results are presented in terms of statistical power to detect the true difference under various scenarios within each missing data mechanism. As noted earlier, sample size was calculated in order to ensure 90% power for data without missing values in each variance-covariance scenario.[14]

[14] With WL variance-covariance, the power was slightly higher than 90% due to rounding-up the required sample size.

5.4.1 Statistical power under MCAR

Figure 5.17 displays the empirical power under MCAR missingness for each of the missing data handling methods under equal and unequal dropout rate between groups. In general, deviation from the nominal power of 90% was increased with respect to increase in overall dropout rate. In LOCF, substantial loss of power was observed when there was equal dropout rate between the groups or higher dropout rate in the experimental group; artificial over-powering was observed with higher dropout in the control group. The loss of power was further influenced by level of data variability. In other methods, missing data led to substantial loss of power in proportion to overall dropout rate; however, the loss was not greatly affected by equal or unequal dropout rate between groups. It can be seen that MMRM had slightly greater power compared to CCA and MI. With 10% dropout rate, the empirical power ranged from 83.5%–87.8% for CCA, from 84.4%–88.1% for MMRM, from 64.0%–93.3% for LOCF, and from 83.3%–87.4%; with 30% dropout rate, it was 73.1%–78.5%, 75.5%–81.1%, 27.9%–95.3% and 71.3%–79.8%, respectively.

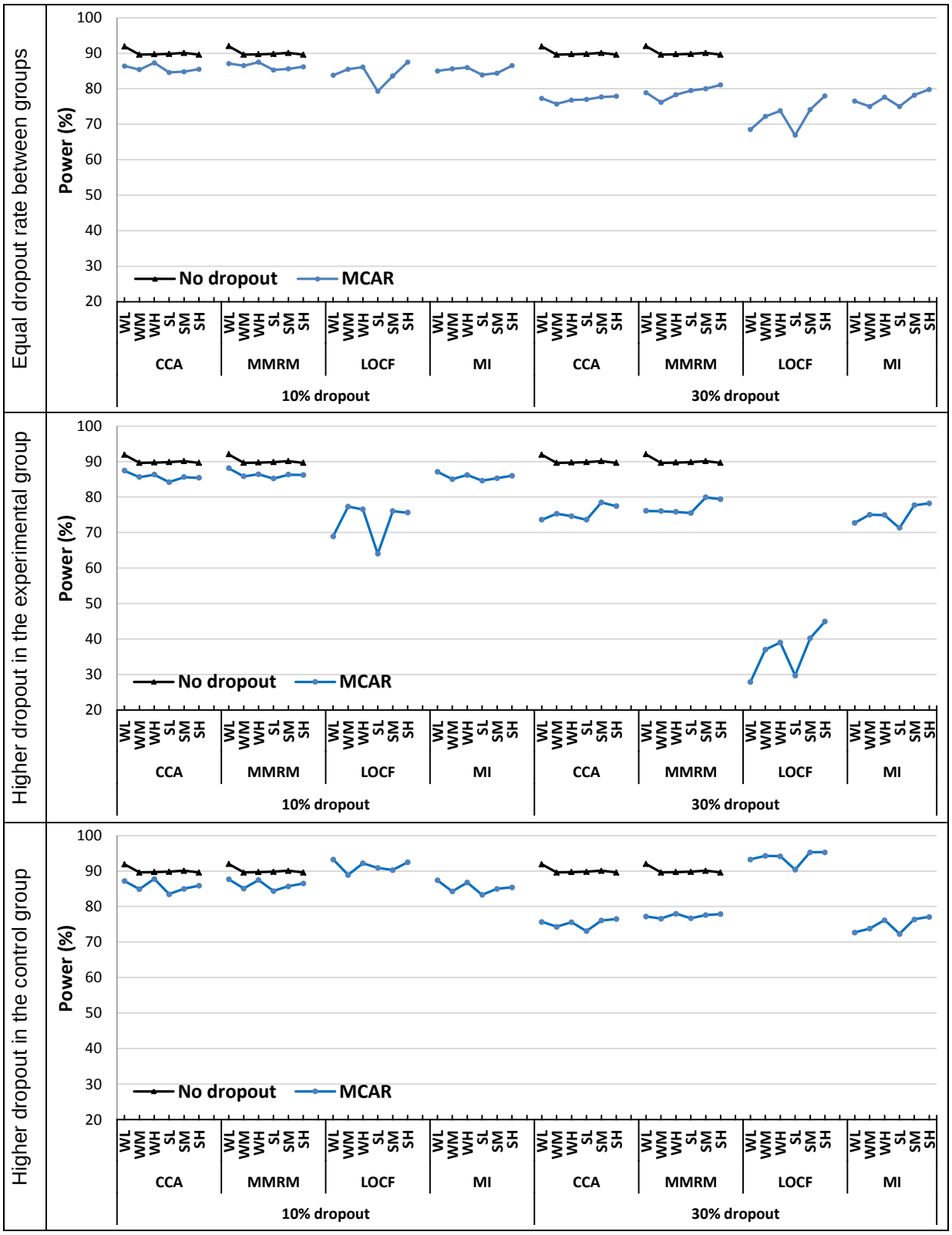


Figure 5.17: Statistical power under MCAR

5.4.2 Statistical power under MAR-B

Figure 5.18 displays the empirical power under MAR-B missingness for each of the missing data handling methods under equal and unequal dropout rate between groups. As in MCAR, LOCF led to substantial deviation from the nominal power in relation to overall dropout rate, dropout rate between groups and direction of dropout. The level of data variability also had an impact on the deviation, especially when the dropouts were in opposite directions between the groups. Similarly, CCA, MMRM and MI also led to loss of power in relation to overall dropout rate, and the loss was slightly higher with differential dropout rates between the groups; but the loss was unaffected by the level of data variability and the correlation between repeated assessments. These methods produced markedly greater power under MAR-B1 compared to MAR-B2. Further, with 30% dropout rate, these methods had greater power under situations where dropout rate was equal between groups compared to unequal dropout. Additionally, MMRM had slightly greater power than MI in most scenarios. With 10% dropout rate and dropouts in the same direction in both groups, the empirical power ranged from 84.3%–88.1% for CCA, from 85.4%–88.8% for MMRM, from 60.4%–95.7% for LOCF, and from 83.1%–87.2% for MI; when dropouts were in opposite directions between the groups, power was 82.5%–87.4%, 84.1%–87.7%, 73.3%–97.9% and 81.3%–86.4%, respectively. With 30% dropout rate and dropouts in the same direction in both groups, the power ranged from 71.2%–79.8% for CCA, from 75.4%–81.2% for MMRM, from 25.5%–96.5% for LOCF, and from 70.7%–79.5% for MI; when dropouts were in opposite directions between the groups, power was 64.9%–74.0%, 68.3%–76.6%, 44.3%–99.6% and 64.3%–74.6%, respectively.

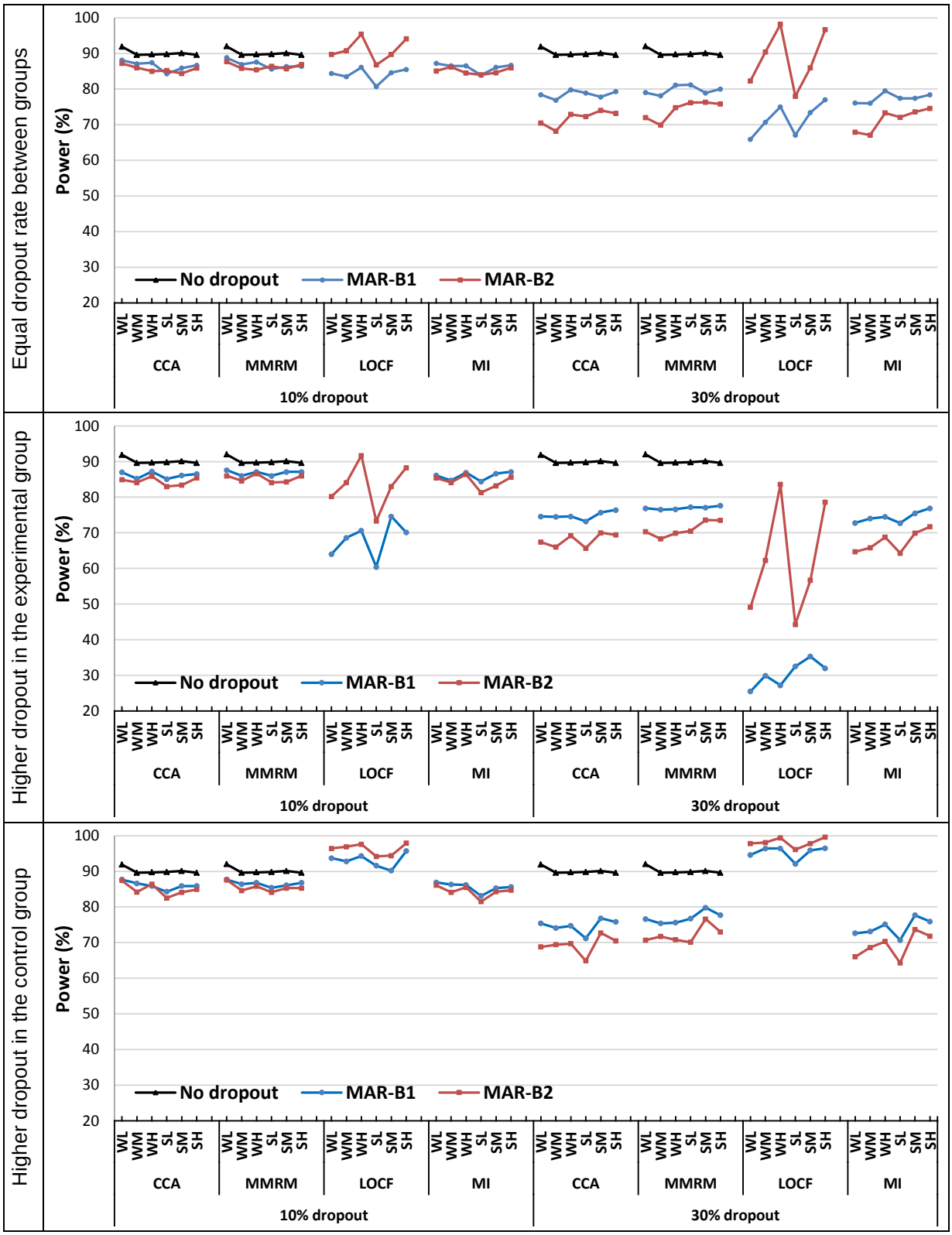


Figure 5.18: Statistical power under MAR-B

5.4.3 Statistical power under MAR-L

Figure 5.19 displays the empirical power under MAR-L missingness for each of the missing data handling methods under equal and unequal dropout rate between groups. Similar to the findings under MCAR and MAR-B, power under LOCF was severely affected under MAR-L – it was either overpowered or underpowered depending upon the direction and magnitude of bias and underestimation of standard error. For example, in the third graph with higher dropout rate in the control group, for data with high SD and weak correlation, the observed power was close to 100% under LOCF when dropouts were in opposite directions between the groups irrespective of whether 10% or 30% dropout rate. In contrast, in the second graph where there was higher dropout rate in the experimental group, for data with high SD and weak correlation the observed power was as low as 5% under LOCF when dropouts were in the same direction in both groups. The deviation was further affected by level of data variability and correlation between repeated assessments. Unlike the situations under MCAR and MAR-B, CCA was also flawed under MAR-L; this approach led to substantial loss of power in most scenarios especially under MAR-L2. In general, for MMRM and MI approaches under MAR-L1, the observed power was nearly similar to that under MAR-B1; however, the estimate under MAR-L2 was markedly higher than that of under MAR-B2. Further, the power of MMRM was noticeably better than that of MI, and was less affected by the direction of dropouts. With 10% dropout rate and dropouts in the same direction in both groups, the empirical power ranged between 79.4%–90.9% for CCA, 85.0%–88.5% for MMRM, 47.6%–97.0% for LOCF, and 82.9%–86.9% for MI; when dropouts were in opposite directions between the groups, it was 68.4%–80.8%, 84.0%–87.6%, 81.0%–99.9% and 82.2%–86.5% respectively. With 30% dropout rate and dropouts in the same direction in both groups, the power ranged between 50.0%–91.2% for CCA, 74.6%–80.7% for MMRM, 4.5%–99.6% for

LOCF, and 70.2%–79.7% for MI; when dropouts were in opposite directions between the groups, it was 25.4%–58.6%, 71.9%–78.7%, 72.8%–100.0% and 64.7%–75.5% respectively.

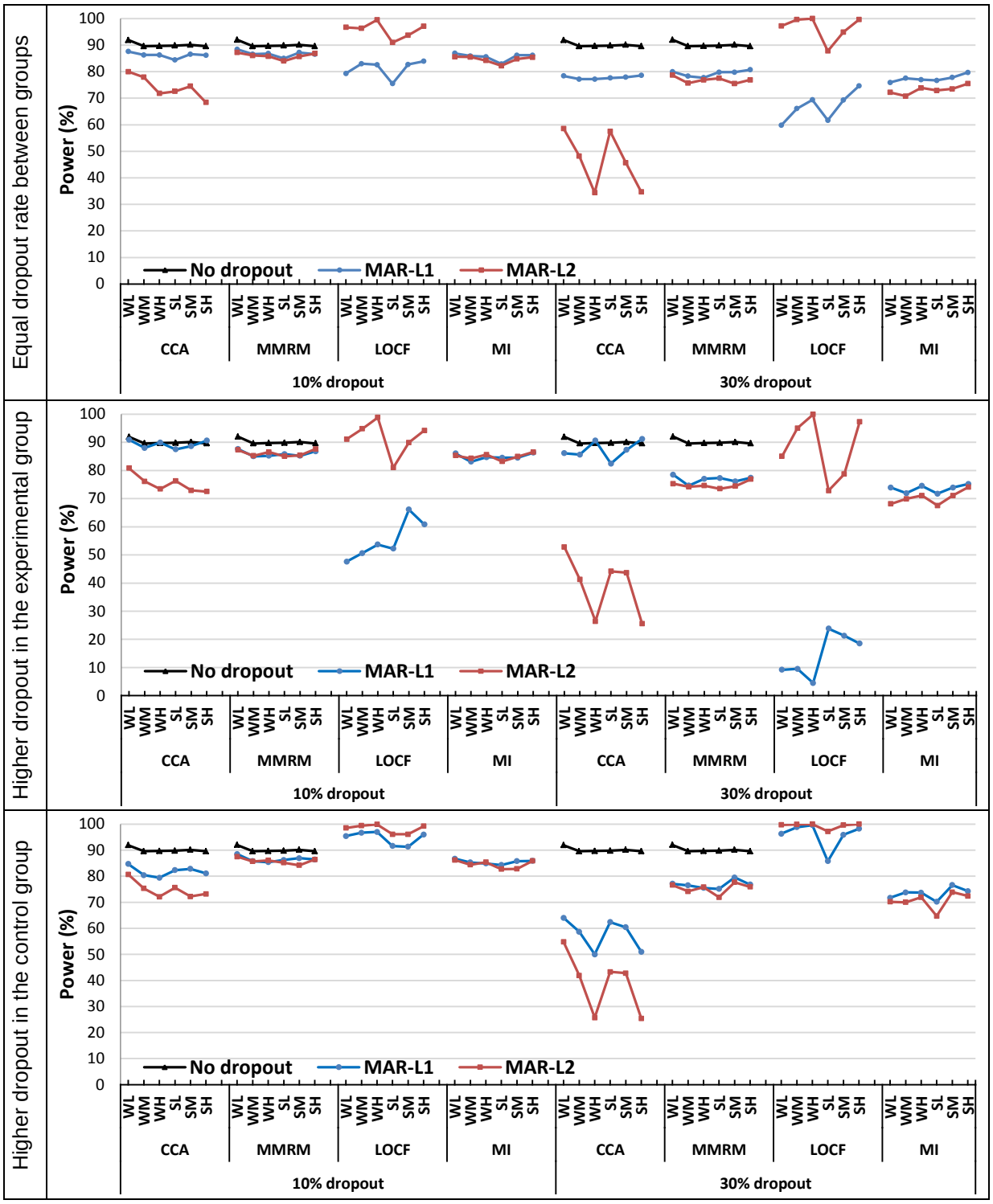


Figure 5.19: Statistical power under MAR-L

5.4.4 Statistical power under MNAR

Figures 5.20a and 5.20b display the empirical power under MNAR for each of the missing data handling methods under equal and unequal dropout rate between groups. Figure 5.20a displays the empirical power in relation to level of dropout rate with a fixed strong correlation and moderate SD. It can be seen that all methods yielded considerably lower power (than the nominal 90% level), as was the case with bias in estimate of treatment effect. Loss of power was substantial even with 10% dropout rate under MNAR-2, and this finding does not differ greatly in respect of equal or unequal dropout rate between groups. Under MNAR-1, observed power for CCA, MMRM and MI was greater than nominal power as the treatment effect was overestimated with these approaches when dropout rate was higher in the experimental group. Figure 5.20b displays the empirical power in relation to data variability and correlation between repeated assessments with 30% dropout rate. It can be seen that all methods were severely flawed, and the deviation from the nominal power was markedly affected by the level of data variability and the correlation in most scenarios, irrespective of equal or unequal dropout rate between groups. Importantly, it can be seen that power was mostly lower than 50% for MNAR-2 scenarios across all analytical approaches – and extending to less than 30% power when dropout was 30% across methods.

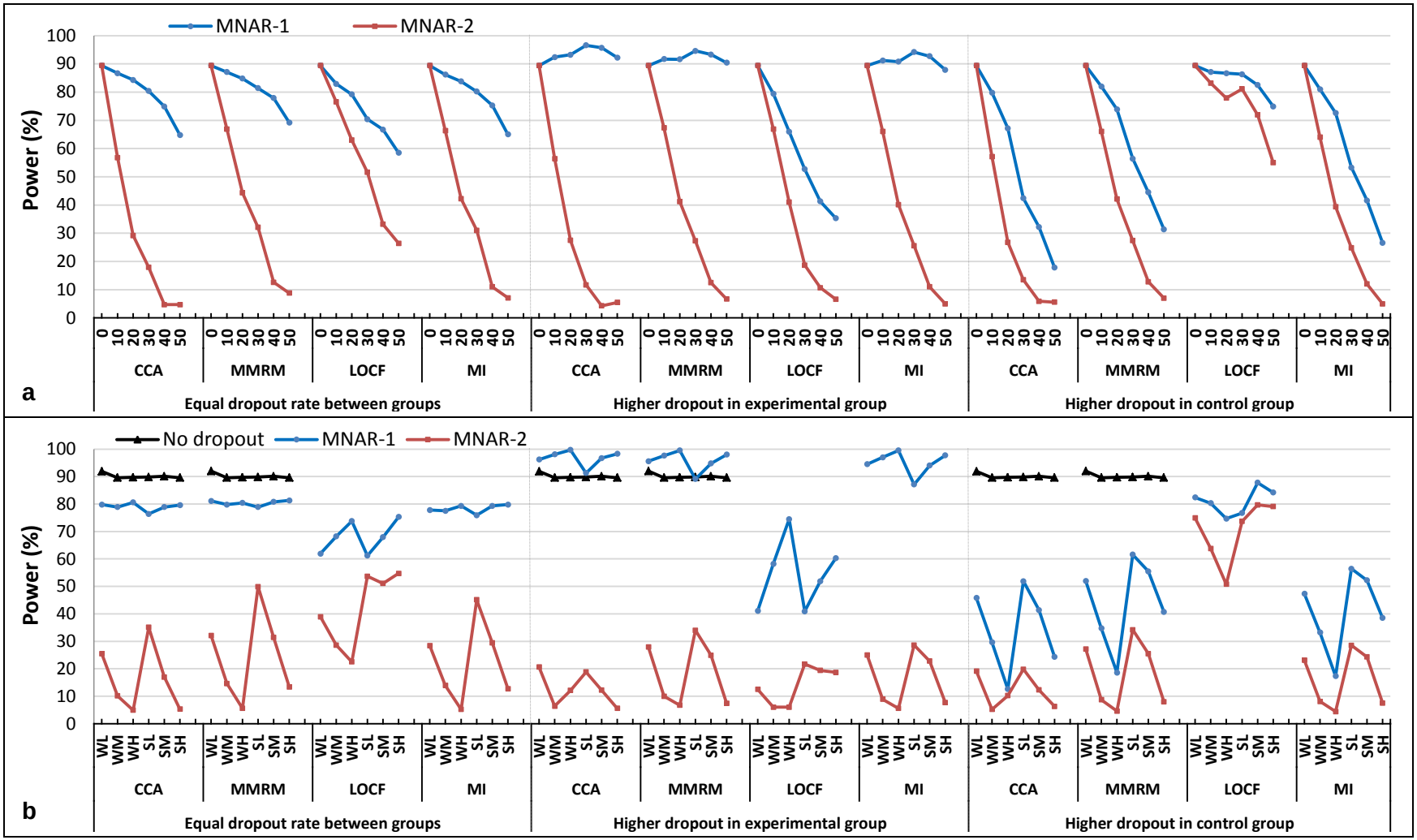


Figure 5.20: Statistical power under MNAR – in relation to (a) level of dropout rate (%) with a fixed strong correlation and moderate SD; (b) data variability and correlation between the repeated assessments with 30% dropout rate

5.5 Summary of findings

Under MCAR, all but LOCF approach yielded an unbiased estimate of treatment effect in all considered scenarios, whereas RMSE increased with an increase in the dropout rate. In this simulation study, LOCF led to overestimation when dropout rate was higher in the control group. Further, larger bias with lower RMSE indicates the underestimation of variability in LOCF. Data characteristics such as spread of data and correlation between repeated assessments did not affect accuracy of CCA, MMRM or MI when the missing data mechanism is MCAR. In addition, all but the LOCF approach retained the targeted CI coverage and width; however, the width in MI was slightly affected by the data variability and the correlation, and was slightly wider than that of CCA and MMRM. LOCF failed to attain the acceptable coverage and leads to underestimation of the width in most scenarios.

Under MAR dependent on baseline value, all approaches but LOCF yielded an unbiased estimate of treatment effect as in the case of MCAR; however, MMRM and MI consistently produced the smallest RMSE, and very slightly lower than that of CCA, compared to LOCF. This study found that the direction of dropouts and level of dropout rate moderately affected the overall accuracy in terms of RMSE, for CCA, MMRM and MI, but particularly in the case of LOCF. Importantly, CCA, MMRM and MI could attain the targeted coverage, even with a 30% dropout rate across all scenarios; MI retained very slightly higher coverage than in MMRM. In these approaches, the CI width was increased in proportion to the overall dropout rate, and was further influenced by equal or differential dropout rates between groups and data variability when the dropout rate was high. It can be found that the data variability had slightly more impact on the width in MI compared to that in MMRM. In addition, MMRM produced a smaller CI compared to CCA and MI. For LOCF, the situation under MAR is worse than for that under MCAR.

Under MAR dependent on last observed value, MMRM and MI yielded unbiased estimates of treatment effect, and consistently produced lower RMSE compared to CCA and LOCF. This study found that the direction of dropouts and level of dropout rate moderately affected the overall accuracy, in terms of RMSE, of MMRM and MI, but substantially for CCA and LOCF. Further, the bias and RMSE of the estimate under CCA and LOCF were influenced by the data variability and the correlation between repeated assessments. Unlike the findings in MAR-B, only MMRM and MI could attain the targeted coverage, irrespective of scenario. CCA led to under-coverage when the dropouts were in opposite directions between the groups; the coverage was markedly reduced in relation to an increase in dropout rate and data variability. With LOCF, the coverage varied substantially across the scenarios; it ranged from 67.2%–97.7% with 10% dropouts and from 6.9%–99.1% with 30% dropouts. It was found that MMRM produced a slightly narrower CI compared to CCA and MI, and the impact of direction of dropout on the width was slightly lower with MMRM compared to the other approaches.

Under MNAR, all approaches were substantially flawed; the bias and RMSE markedly increased in relation to an increase in overall dropout rate and data variability. However, with equal dropout rate between groups, the estimates of treatment effect in CCA, MMRM and MI were less adversely affected by the dropout rate when the dropouts were in the same direction. Further, both MMRM and MI yielded similar estimates of the treatment effect and RMSE in all scenarios; these measures were more appropriate than those given by CCA, especially when the correlation was strong. In this simulation under an MNAR mechanism, it has been noted that LOCF was the most favourable approach compared to others in few circumstances; however, mean change over time had significant effect on LOCF – the effect of mean trajectory profile is reported in chapter 6. Since the bias in estimate of treatment effect was a major problem with all the approaches, the coverage and power were also substantially deviated from the targeted level.

Chapter 6: Simulation study: findings 2

6.1 Introduction

This chapter presents the results of simulation studies 2, 3 and 4 (detailed in chapter 4) that were designed to examine the relative performance of four missing data handling approaches – CCA, MMRM, LOCF and MI – when analysing incomplete longitudinal RCT data under the four missing data mechanisms: MCAR, MAR-B, MAR-L and MNAR. Analysis of covariance (ANCOVA) was used in conjunction with CCA, LOCF and MI. As proposed in the introduction chapter, the goal was to answer the following research questions:

i. Within and across the missing data handling approaches, do the accuracy and efficiency of the parameter estimates (i.e. treatment effect at the primary endpoint) vary by changes in an average trajectory pattern over a study period and by size of the treatment effect at the endpoint, given a fixed variance-covariance matrix under the missing data mechanisms?

ii. Does considering baseline of an outcome measure as part of an outcome vector in MMRM analysis have an advantage over the analysis with baseline-as-covariate when there are participants without follow-up data?

iii. Does an increment in sample size in proportion to an expected dropout rate help to achieve the desired statistical power when using the missing data handling approaches under the missing data mechanisms?

6.2 Effect of trajectory pattern and size of treatment effect on inferences from the missing data handling approaches

This study was planned in order to evaluate the effect of mean trajectories with the same treatment effect at the primary endpoint (trajectories 1 and 2 with the same treatment effect of −9.0), and the effect of mean trajectories with different treatment effects at the endpoint (trajectories 1, 3 and 4 with treatment effects of −9.0, 0 [no effect] and −18.0, respectively). Trajectory 1 assumed both treatments improved over time. However, the treatment group improved more and thus resulted in a treatment benefit over time. In contrast to trajectory 1, trajectory 2 assumed that the control group showed better improvement during the initial visit (visit 1) than the experimental group but it retained the same treatment effect at the end as in trajectory 1. Trajectory 3 assumed both treatments improved equally well, and reflected the null hypothesis that there was no difference between treatments at the primary endpoint. Trajectory 4 assumed the experimental group improved quickly but then showed little change over time; however, there was minor improvement in the control group and a treatment effect of −18.0 at the primary endpoint was assumed in favour of the experimental group. This study compared bias, RMSE, and width and coverage of 95% CI in respect of the estimate of treatment effect from the missing data handling approaches under various dropout scenarios for a given SM covariance (strong correlation and moderate SD) matrix, sample size (n = 60 per group) and 30% dropouts.

Figures 6.1–6.4 display the bias, RMSE, and coverage and width of 95% CI respectively. As expected, the performance measurements in respect of the estimate of treatment effect from CCA, MMRM and MI were similar across the mean trajectories (i.e., independent of the trajectory pattern and the size of treatment effect) under the various dropout scenarios. The CCA performed well under the missing data mechanisms: MCAR, MAR-B1, MAR-B2, MAR-L1 (with equal dropout rate between arms) and MNAR-1 (with equal dropout rate between arms). In addition to these mechanisms, the MMRM and MI ANCOVA analyses also performed well under MAR-L1 (with differential dropout rates between arms) and MAR-L2. These results also confirmed the findings in chapter 5.

By contrast, the performance of LOCF varied considerably across the trajectories, especially under a scenario where there was higher dropout rate in the experimental group than in the control group. The considerable variation in this scenario was due to the steep reduction in outcome score over time in the experimental group compared to the control group for trajectories 1 and 2. Despite a high treatment effect at the primary endpoint, LOCF performed relatively better with trajectory 4 than with trajectories 1 and 2 because a high treatment effect was attained so quickly and remains constant over time. It was clear from these results that the standard classification of the missing data mechanism is irrelevant to justify the use of LOCF imputation.

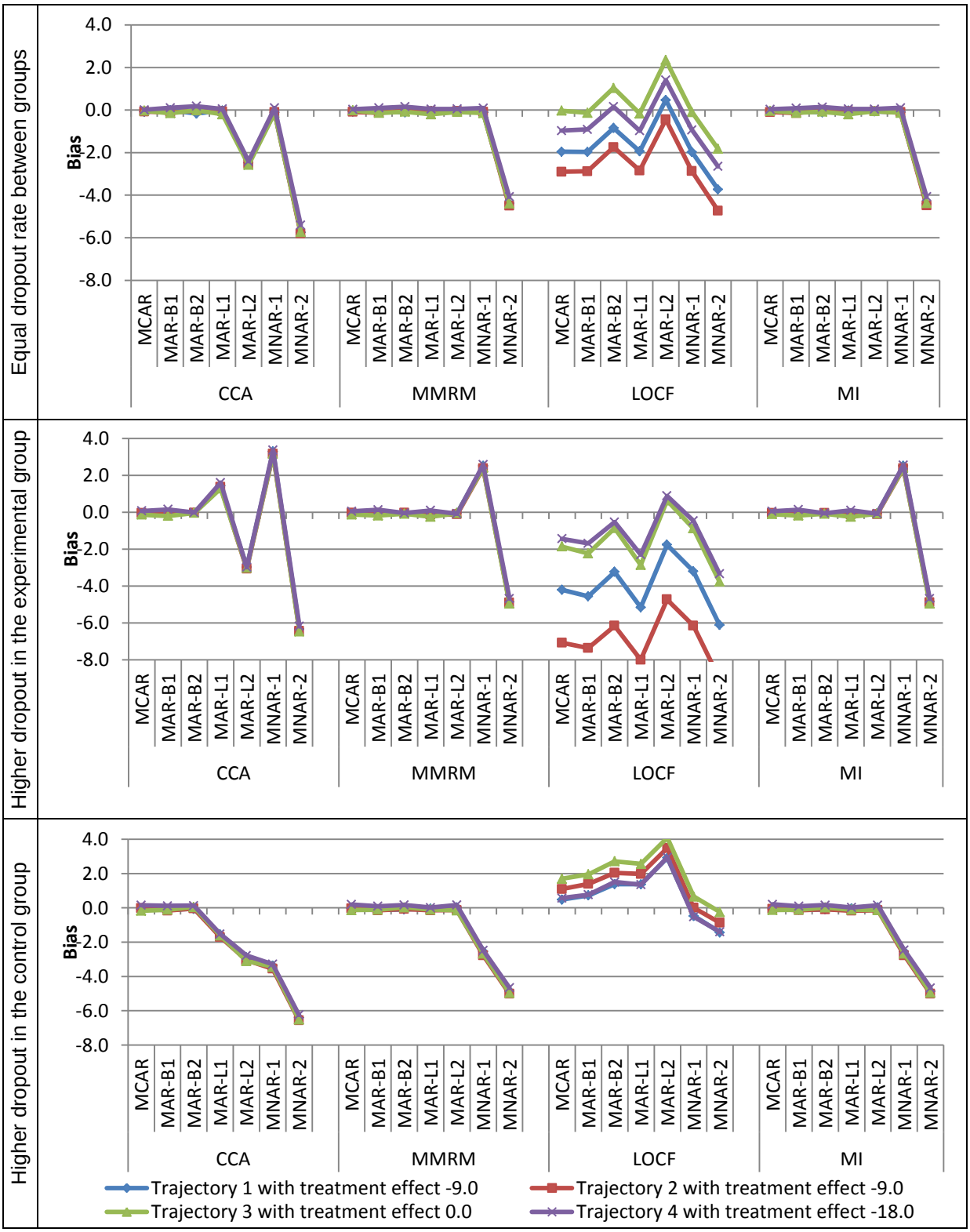


Figure 6.1: Effect of trajectory pattern and mean difference between groups over time on bias in estimate of treatment effect

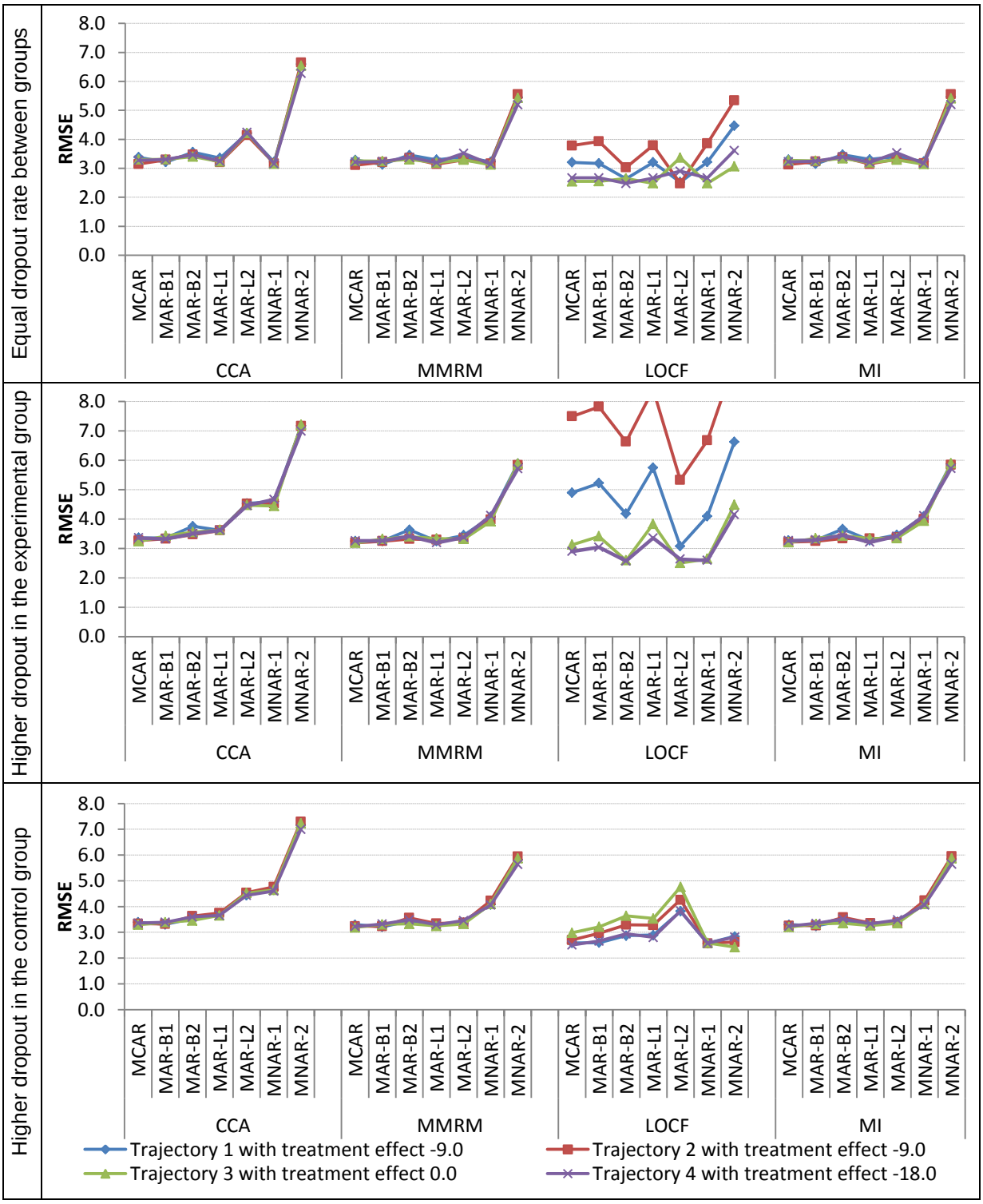


Figure 6.2: Effect of trajectory pattern and mean difference between groups over time on RMSE of estimate

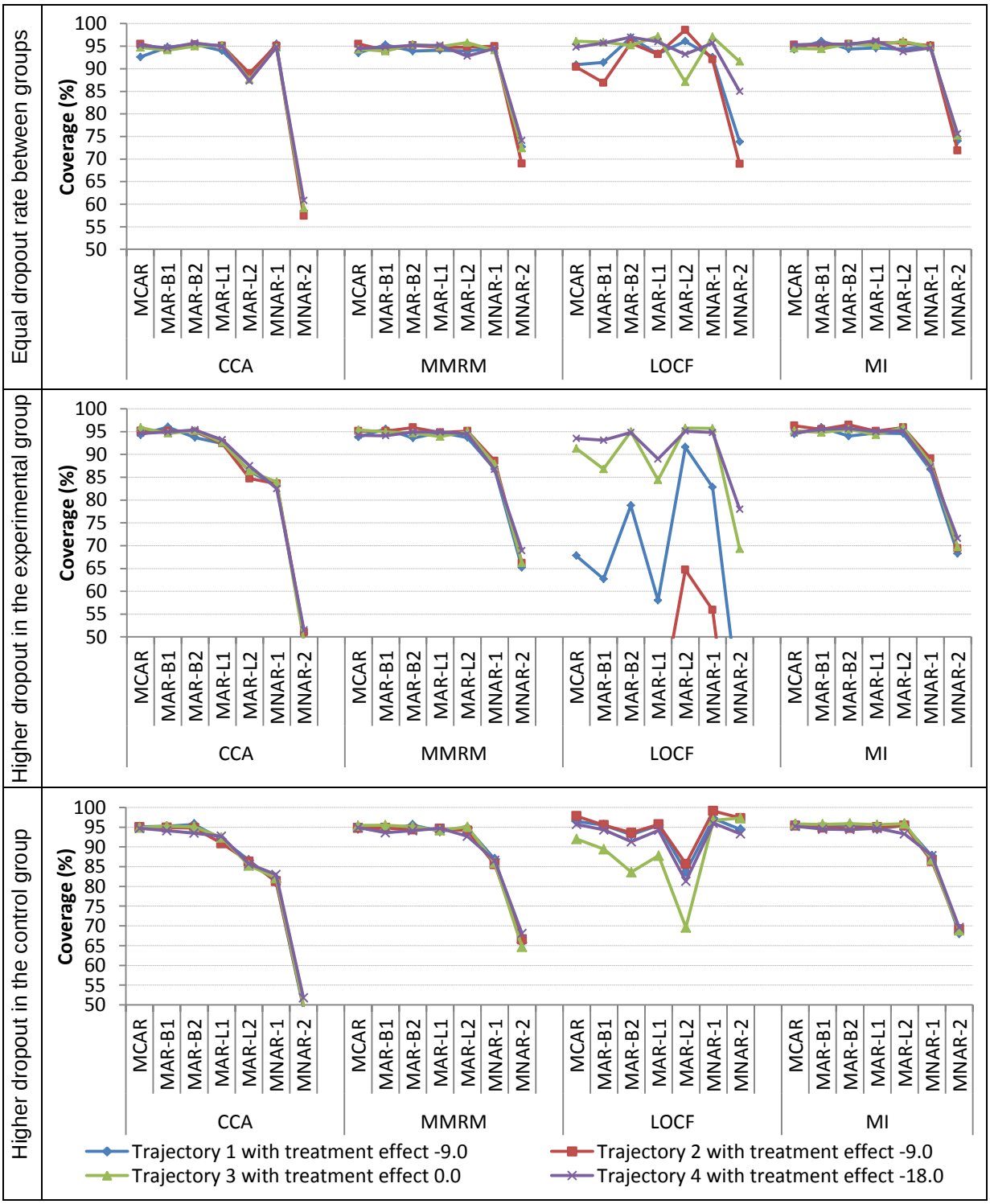


Figure 6.3: Effect of trajectory pattern and mean difference between groups over time on coverage of 95% CI

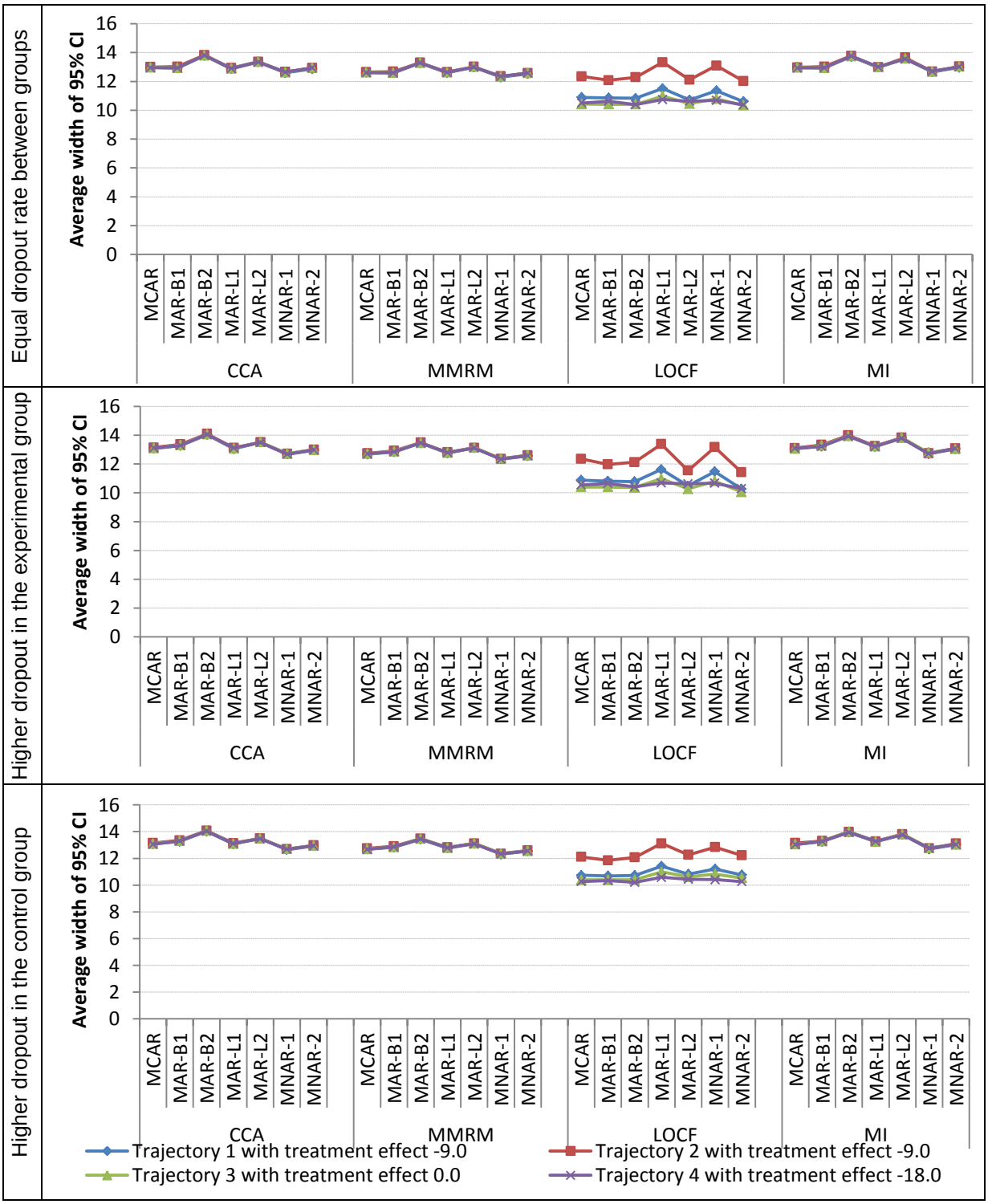


Figure 6.4: Effect of trajectory pattern and mean difference between groups over time on width of 95% CI

6.3 Comparison of two strategies for handling baseline data in an MMRM analysis

This section shows the findings for comparison of two strategies for handling baseline observations in a mixed-effects model for repeated measures analysis under various dropout scenarios. In the first strategy (denoted as MMRM), the baseline score of a repeated outcome measure was included as a covariate in the analysis of post-randomization outcome data in a repeated measures analysis model. In order to obtain baseline-adjusted estimates, baseline–follow-up visit interaction was specified in the analysis model. This model was used in all other simulation studies in this thesis. In the second strategy (denoted as constrained longitudinal data analysis [cLDA]), the baseline was included together with the post-randomization outcome measurements in the context of the outcome variable. Although the outcome variable in the model includes the baseline measures, a constrained term is added to the model to specify that the true baseline means are the same for different treatment groups due to randomization, and this analysis provides baseline-adjusted estimate of treatment effect. Both strategies were described in chapter 2. A detailed discussion around the simulation methodology for this comparison was provided in chapter 4.

There is subtle contrast between the two strategies in the number of participants that are included in the analysis. The analysis with baseline-as-outcome (strategy 2; cLDA) includes all participants with either baseline or follow-up outcome data. By contrast, the analysis with baseline-as-covariate (strategy 1; MMRM) includes only those participants who provided outcome data at baseline and at least one follow-up. Since the simulated datasets in this thesis involve early dropouts without any follow-up assessments, those dropouts were excluded from the MMRM analysis. Therefore, cLDA was used to evaluate the robustness of the analysis results to those exclusions under different missing data mechanisms. All

analyses were performed with restricted maximum likelihood instead of maximum likelihood estimation, but with and without the Kenward-Roger correction for finite sample. Since there is no finite sample correction implemented in conjunction with Stata *xtmixed* procedure (StataCorp, 2011), this correction was performed using an add-on within the SAS *proc mixed* procedure (SAS Institute Inc., 2011). Without the Kenward-Roger correction, SAS and Stata produced identical results if the parameter estimation in Stata *mixed* was set to be based on a t-distribution (the customary default option in Stata is that of a standard normal distribution). Data simulation was performed with the SM covariance (strong correlation and moderate SD) matrix and 30% overall dropout rate – nearly 10% of participants had not completed any follow-up – under different missing data mechanisms.

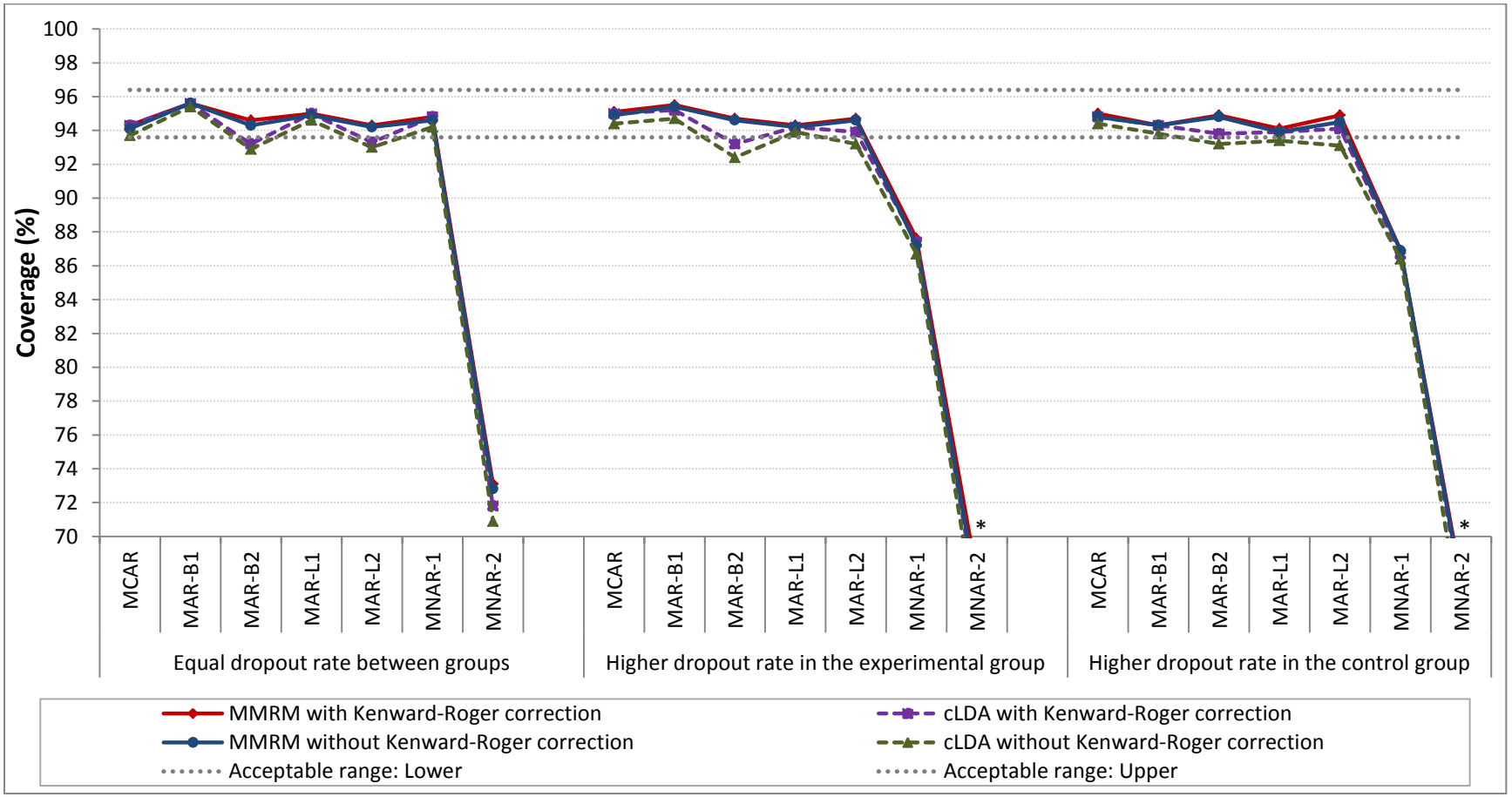
6.3.1 Effects on overall accuracy

The overall accuracy – in terms of bias and RMSE – of the estimate of treatment effect was identical across the strategies in all the scenarios considered in this simulation. These results are provided in appendix 6 (Tables 23 and 24). However, the study found slight differences in average SE between these strategies, and this therefore influenced the coverage of the 95% CI and statistical power.

6.3.2 Effects on the coverage of 95% CI

Figure 6.5 displays the 95% CI coverage for each of the strategies under equal and unequal dropout rate between groups. The red and blue solid lines represent the observed coverage with MMRM with and without Kenward-Roger correction, respectively. The purple and green dashed lines represent the corresponding observed coverage with cLDA.

It was found that the MMRM analysis yielded a very similar coverage probability with and without Kenward-Roger correction. However, the correction led to a slightly better coverage (i.e. coverage closer to 95%) in the cLDA, irrespective of the dropout scenarios considered in this study. When comparing the baseline handling strategies, the MMRM analysis yielded very slightly better coverage compared to the cLDA, especially when the dropouts were in opposite directions between the intervention groups. Under MAR-B2 with equal dropout rate between the groups, the observed coverage increased from 92.9% (without Kenward-Roger correction) to 94.3% by considering baseline as a covariate rather than an outcome. Under MAR-L2, it increased from 93.0% to 94.2%. Therefore, as seen in the figure, the coverage with MMRM (i.e. model with baseline-as-covariate) was unlikely to be affected by the direction of dropouts in scenarios where the estimate of treatment effect was unbiased – similar results were found in chapter 5 (section 5.3).



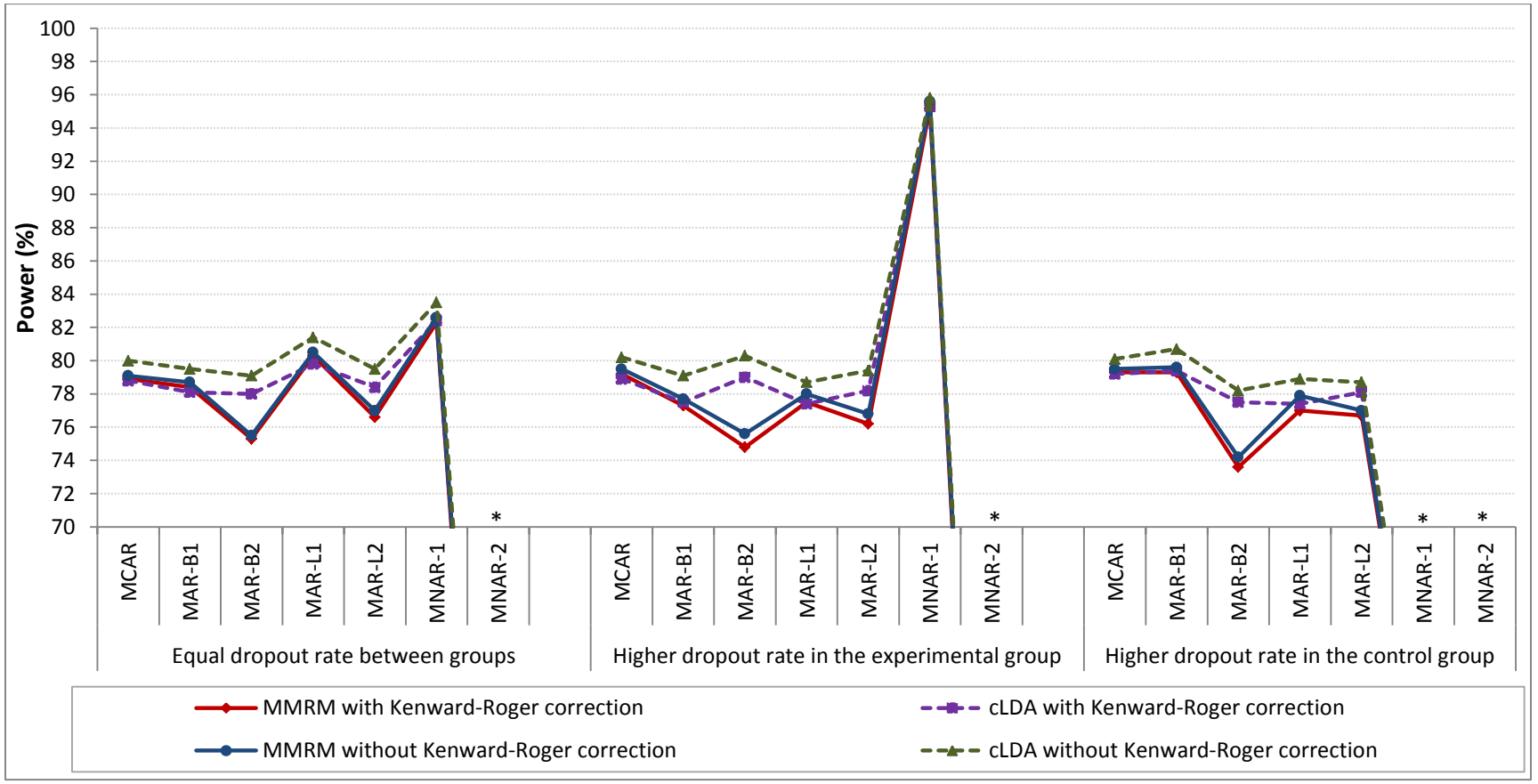
*Coverage was less than 70%

Figure 6.5: Coverage of 95% CI under various scenarios (MMRM – Mixed model with baseline-as-covariate; cLDA – Mixed model with baseline-as-outcome)

6.3.3 Effects on the observed power

Figure 6.6 displays the observed statistical power for each of the strategies under equal and differential dropout rate between groups. The effect of the Kenward-Roger correction on power differed from the effect on coverage. That is, due to the correction, power was decreased slightly with cLDA but not that much with MMRM. It appeared that cLDA without Kenward-Roger correction (green dashed line) yielded the highest power, and MMRM with Kenward-Roger correction (red solid line) yielded the lowest power among the strategies to handle baseline data.

As expected, due to a higher number of subjects included in the analysis, cLDA yielded a higher power compared to MMRM, and the power with cLDA was less likely to be affected by the direction of dropouts. Specifically, under MCAR and MAR with dropouts in the same direction in both groups (i.e. MAR-B1 and MAR-L1), the empirical power was in the range of 78%–80% across all strategies except cLDA without the correction. Under MAR with dropouts in opposite directions between groups (i.e. MAR-B2 and MAR-L2), the power of cLDA with the correction was similarly ranged at about 80% as in MAR-B1 and MAR-L1; whereas the power of MMRM with/without the correction ranged between 74%–77%. Under MNAR, the power comparison was not relevant as all strategies were flawed with substantial bias in the estimate of treatment effect.



*Power was less than 70%

Figure 6.6: Statistical power under various scenarios (MMRM – Mixed model with baseline-as-covariate; cLDA – Mixed model with baseline-as-outcome)

6.4 Effect of sample size on power under different missing data mechanisms: a comparison of missing data handling approaches

The effect of missing data on the statistical power of clinical trials has been detailed in chapter 5. The simulation studies have established that trials could be artificially underpowered/overpowered depending on the magnitude and direction of bias in estimates. It was also found that trials are still underpowered even with unbiased estimates of treatment effect irrespective of analysis methods or missing data mechanisms. When there is no bias in the estimate of treatment effect, the reduction of power was shown to be associated with the direction of dropout in addition to the amount of missing data. The reduction was relatively larger in situations where dropouts were in opposite directions between study groups compared to the situation where dropouts were in the same direction in both groups.

The current practice in the presence of anticipated missing data is simply to inflate the sample size that was calculated assuming no missing data, based on the inverse of one minus the anticipated dropout proportion. Further simulation studies were performed here to verify how far this current practice in sample size calculation protects against the loss of statistical power due to missing data. With a WM[15] variance-covariance matrix among the repeated measurements, and 10% dropout rate, these studies used two sample sizes: n = 150 (without inflation for the missing data) – this sample size ensured 90% power to detect the true treatment effect at an endpoint in the absence of missing data; and n = 168 (inflated for 10% dropouts). These studies were repeated for the 30% dropout scenario with n = 150 (without inflation for the missing data) and n = 216 (inflated for 30% dropouts). The

[15] Weak correlation and moderate SD

corresponding scenarios with 80% power were also explored: an unadjusted sample size of 57 per group was used. The corresponding inflated sample sizes were 63 per group and 81 per group when adjusted for 10% and 30% dropout rates, respectively. Unlike the studies in sections 6.2 and 6.3, this simulation used a weak correlation matrix because the earlier simulation studies observed lower power for weak correlation scenario compared to the strong correlation.

6.4.1 When the desired power was 90% in the absence of missing data

As expected, all approaches but LOCF yielded similar estimates of treatment effect and coverage probability across the sample sizes (results not shown), which was equal to the estimates found in chapter 5 under corresponding missing data scenarios. Figures 6.7 and 6.8 display the observed statistical power – when the desired power was 90% in the absence of missing data – for each analytical method in relation to the sample sizes under 10% and 30% dropout scenarios, respectively. The figures are split in order to show results stratified according to balance/imbalance in dropout rate between treatment groups: (i) equal dropout rate between the groups; (ii) higher dropout rate in the experimental group; and (iii) higher dropout rate in the control group. The dashed line indicates the observed power corresponding to scenarios where no attrition adjustment to the required sample size was made. The solid line indicates the observed power corresponding to scenarios where the required sample size was inflated. Table 6.1 shows the observed power for CCA, MMRM and MI for inflated sample sizes of 10% and 30% dropout rates.

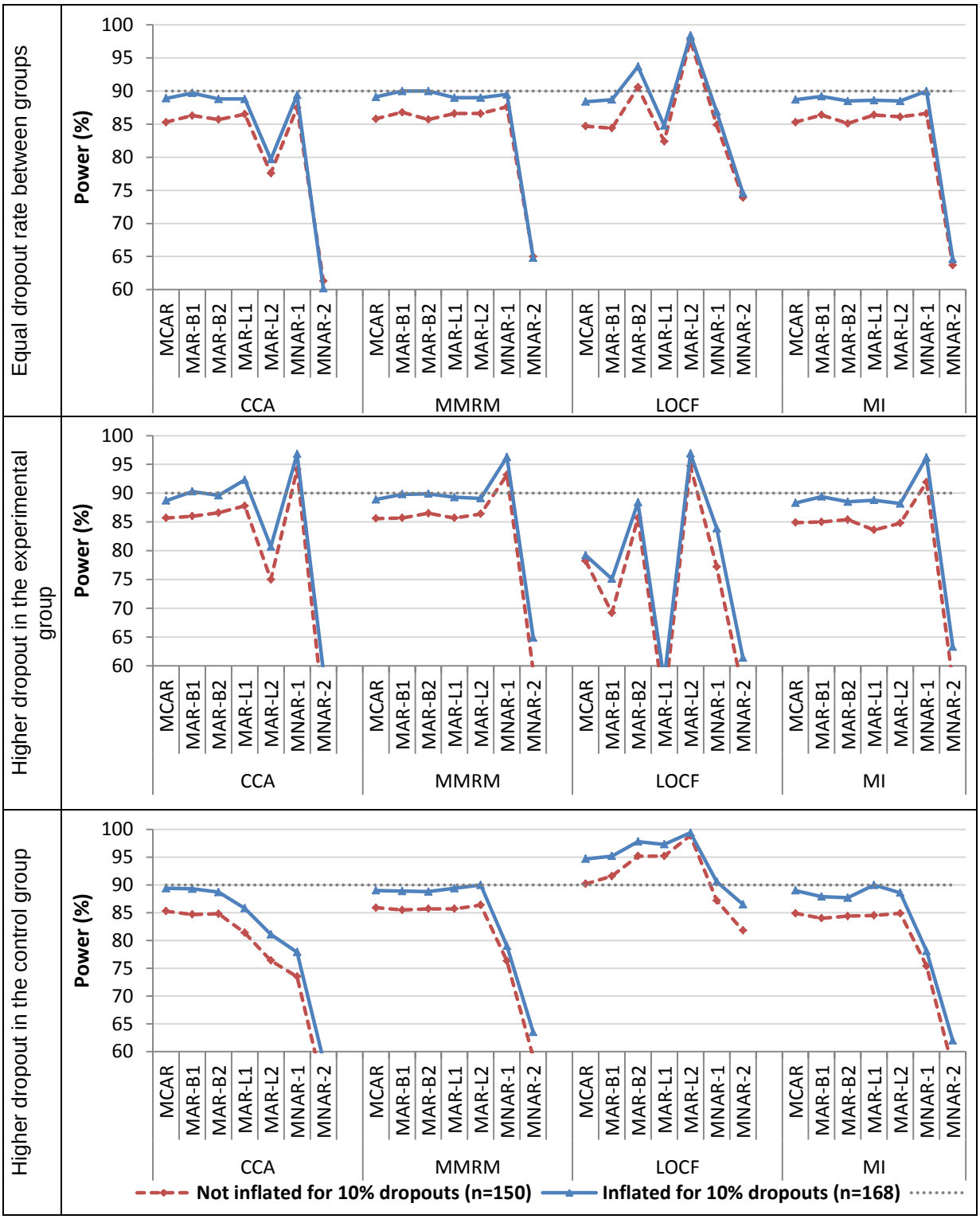


Figure 6.7: Statistical power under different sample sizes (10% dropouts)

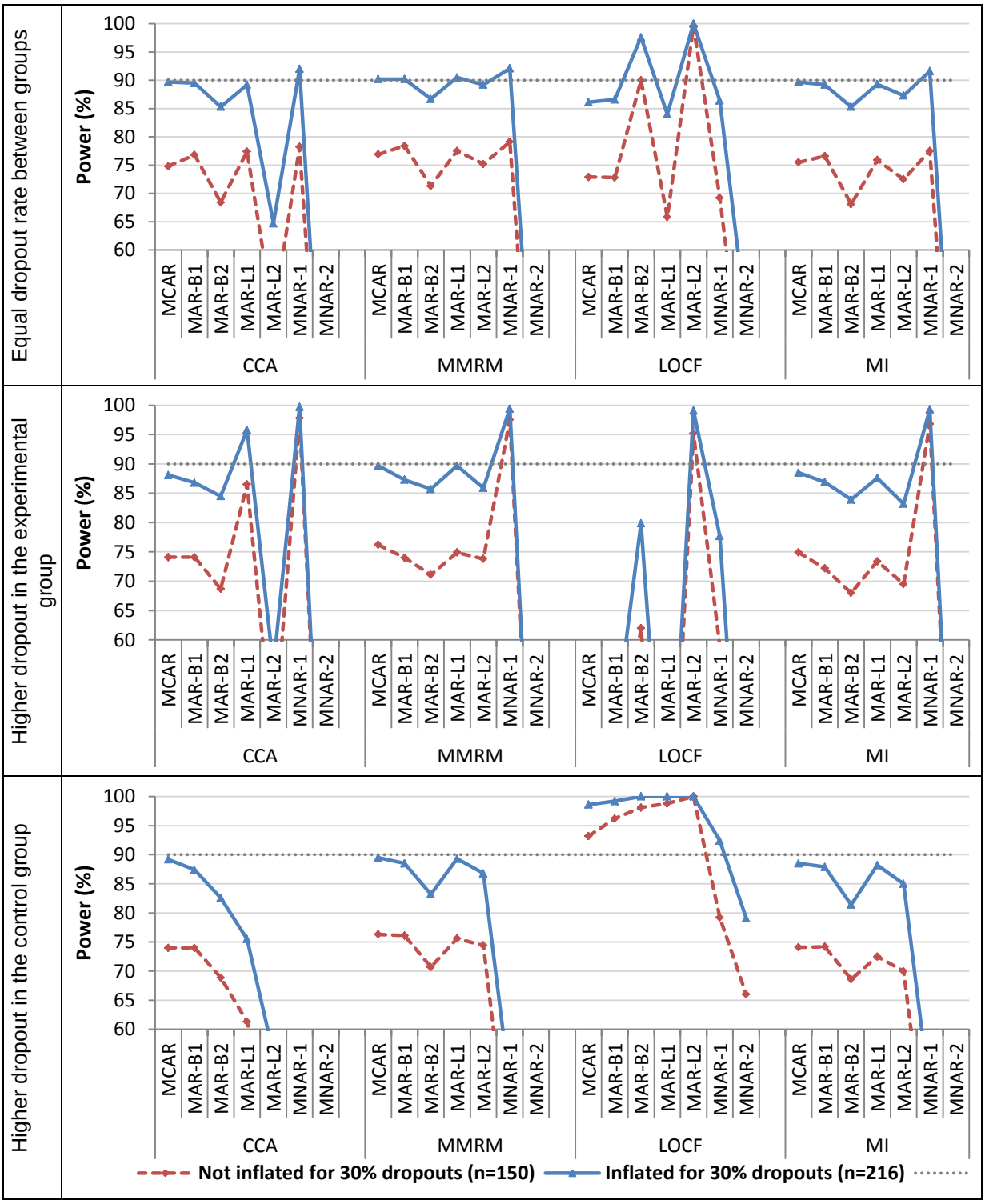


Figure 6.8: Statistical power under different sample sizes (30% dropouts)

Table 6.1: The observed power (%) with inflated sample size – desired power was 90%

Mechanism	10% dropouts (n = 168)			30% dropouts (n = 216)		
	CCA	MMRM	MI	CCA	MMRM	MI
Equal dropout between groups						
MCAR	88.9	89.1	88.7	89.7	90.2	89.7
MAR-B1	89.7	90.0	89.2	89.5	90.2	89.2
MAR-B2	88.8	90.0	88.5	85.3	86.7	85.3
MAR-L1	88.8	89.0	88.6	89.2	90.5	89.3
MAR-L2	79.7*	89.0	88.5	64.7*	89.2	87.3
MNAR-1	89.4	89.5	90.0	92.0	92.1	91.6
MNAR-2	60.2*	64.8*	64.6*	15.1*	22.9*	21.5*
Higher dropout in the experimental group						
MCAR	88.7	88.9	88.3	88.1	89.7	88.5
MAR-B1	90.3	89.8	89.4	86.8	87.3	86.9
MAR-B2	89.6	89.9	88.5	84.5	85.7	83.9
MAR-L1	92.3*	89.3	88.8	95.8*	89.7	87.6
MAR-L2	80.7*	89.1	88.2	55.3*	85.9	83.2
MNAR-1	96.8*	96.3*	96.2*	99.7*	99.4*	99.3*
MNAR-2	59.2*	64.9*	63.3*	8.3*	13.5*	12.4*
Higher dropout in the control group						
MCAR	89.4	89.0	89.0	89.2	89.5	88.5
MAR-B1	89.3	88.9	87.9	87.4	88.5	87.9
MAR-B2	88.7	88.8	87.7	82.6	83.2	81.4
MAR-L1	85.8*	89.4	90.0	75.5*	89.3	88.2
MAR-L2	81.1*	90.0	88.6	55.7*	86.8	85.0
MNAR-1	77.9*	79.0*	78.1*	41.3*	50.2*	48.9*
MNAR-2	58.6*	63.5*	62.0*	8.4*	14.0*	14.0*

*Estimates of treatment effect were biased

As seen in chapter 5 (section 5.4), the bias in the estimate of treatment effect caused underestimation or overestimation of the power to detect the true treatment effect and was dependent on the direction and size of the bias. In figures 6.7 and 6.8, the observed power was slightly higher with MMRM in comparison with CCA and MI in scenarios where the estimate of treatment effect was unbiased, and the power of MI was no better than CCA in these scenarios. The same observations were made in chapter 5 (section 5.4) in relation to the WM covariance matrix. Those studies in chapter 5 further showed that MI could perform well in comparison to CCA in terms of the observed power when the baseline-endpoint correlation increased.

As seen in figures 6.7 and 6.8, the observed power in all methods was substantially increased in relation to increases in sample size. Therefore, the loss of power due to attrition was substantially reduced with the increased sample size in scenarios where the analytical methods – CCA, MMRM and MI – provide unbiased estimates of treatment effect (Table 6.1). Importantly, in these scenarios, CCA, MMRM and MI could attain the observed power close to the nominal level of 90% by using the increased sample size when dropout rate was equal and dropouts were in the same direction in both groups, irrespective of level of dropout rate (10% or 30%) and dropout mechanism (MCAR, MAR or MNAR).

With 10% dropout rate and the inflated sample size (n = 168), neither the differential dropout rates between the groups or the direction of dropouts prevented the desired power of 90% from being attained (Figure 6.7 and Table 6.1). Importantly, in scenarios where the estimate of treatment effect was unbiased, the observed power with the inflated sample size was higher than 88.7% for CCA, higher than 88.8% for MMRM and higher than 87.7% for MI. The lowest power among these methods was observed under an MAR-B2 mechanism.

With 30% dropout rate (Figure 6.8 and Table 6.1), the inflated sample size maintained the desired power in CCA, MMRM and MI under an MCAR mechanism. Under MAR, the observed power in these methods was slightly lower than the desired level for the case of differential dropout rate between the groups, but the loss was substantial when dropouts were in opposite directions between the groups (MAR-B2 and MAR-L2). With increased sample size (n = 216), the observed power among the methods with unbiased estimate of treatment effect was as low as: 82.6% for CCA, 83.2% for MMRM, and 81.4% for MI. As in the case of a 10% dropout rate, the lowest power for these methods was observed under an MAR-B2 mechanism. The observed power for MMRM and MI under an MAR-L mechanism was better than the corresponding power under MAR-B2, and even with differential dropout scenarios: the observed power ranged 85.9%–90.5% for MMRM and 83.2%–89.3% for MI.

Similar findings were observed when the simulation study was repeated with an SW variance-covariance (strong correlation and moderate SD) matrix (Appendix 7: table 25).

6.4.2 When the desired power was 80% in the absence of missing data

Table 6.2 provides the observed statistical power – when the desired power was 80% in the absence of missing data – for CCA, MMRM and MI in relation to the inflated sample sizes under 10% and 30% dropout scenarios. Graphical presentations of these results are given in appendix 7 (Figures 1 and 2). The findings were similar to the scenarios where the nominal power was 90% in the absence of missing data.

Table 6.2: The observed power (%) with inflated sample size – desired power was 80%

Mechanism	10% dropouts (n = 114)			30% dropouts (n = 162)		
	CCA	MMRM	MI	CCA	MMRM	MI
Equal dropout between groups						
MCAR	78.9	79.2	78.0	81.0	82.1	80.0
MAR-B1	78.8	79.5	78.4	81.0	82.1	80.2
MAR-B2	77.7	78.8	77.0	75.5	76.6	73.6
MAR-L1	81.1	80.6	79.5	80.3	81.2	79.4
MAR-L2	69.0*	80.9	80.0	53.1*	80.6	78.6
MNAR-1	81.2	81.4	80.0	81.5	82.1	81.1
MNAR-2	48.1*	54.9*	52.7*	13.2*	18.6*	17.5*
Higher dropout in the experimental group						
MCAR	80.1	80.7	79.8	79.0	80.1	78.5
MAR-B1	80.0	80.0	78.3	79.2	80.3	77.4
MAR-B2	78.4	78.6	78.3	72.7	73.9	71.5
MAR-L1	82.3*	80.1	79.1	89.1*	79.2	77.4
MAR-L2	68.8*	78.1	76.9	47.5*	78.2	74.1
MNAR-1	89.3*	88.8*	87.4*	98.8*	97.9*	98.0*
MNAR-2	48.3*	53.3*	51.7*	8.7*	15.0*	13.6*
Higher dropout in the control group						
MCAR	80.1	80.9	79.3	80.3	81.1	79.9
MAR-B1	79.3	79.4	78.3	77.4	78.7	76.8
MAR-B2	77.6	78.6	77.9	74.7	76.9	73.9
MAR-L1	76.5*	80.0	78.8	65.3*	80.6	79.3
MAR-L2	69.6*	78.9	78.0	48.0*	81.4	77.3
MNAR-1	71.8*	73.7*	72.4*	35.6*	42.1*	39.6*
MNAR-2	48.7*	53.7*	52.3*	8.8*	14.8*	14.5*

*Estimates of treatment effect were biased

With 10% overall dropout rate (Table 6.2), the inflated sample size retained the observed power closer to the desired power in those methods considered here when the estimate of treatment effect was unbiased. The observed power did not vary greatly with differential dropout rate or the direction of dropouts. In scenarios where the estimate of treatment effect was unbiased, the observed power with the inflated sample size was higher than 77.6% (under MAR-B2 with higher dropout in the control group) for CCA, higher than 78.1% (under MAR-L2 with higher dropout in the experimental group) for MMRM, and higher than 76.9% (under MAR-L2 with higher dropout in the experimental group) for MI.

Similar findings were observed when the simulation study was repeated with an SW covariance (strong correlation and moderate SD) matrix (Appendix 7: table 26).

With 30% dropout rate (Table 6.2), the inflated sample size maintained the desired power in CCA, MMRM and MI under an MCAR mechanism. Although the estimate of treatment effect was unbiased, CCA failed to attain the nominal power with the inflated sample size under MAR-B2. The loss was substantial when dropout rate was higher in the experimental group – the observed power was 72.7%. In all missing data mechanisms except MAR-B2, where the estimate of treatment effect was unbiased, MMRM demonstrated an observed power closer to the desired level. In MAR-B2, the observed power with MMRM was as low as 73.9% in the scenario with higher dropout rate in the experimental group; however, slightly higher power was observed under a scenario with an equal dropout rate between groups (the observed power was 76.6%) or higher dropout rate in the control group (the observed power was 74.7%). MI also demonstrated an observed power closer to the desired level after increasing the sample size under scenarios where dropout rate was equal between groups, except for MAR-B2 and MNAR-2 missing data mechanisms. The loss of statistical power in MI after increasing the sample size was noticeable compared to MMRM in scenarios where

differential dropout occurred, and the loss of power was substantial in scenarios where dropouts were in opposite directions between the groups. In scenarios where the estimate of treatment effect was unbiased, the observed power of MI was considerably lower under MAR-B2 (73.6% with equal dropout rate between groups, 71.5% with higher dropout in the experimental group, and 73.9% with higher dropout in the control group) than under MAR-L2.

6.5 Summary of findings

In the first part of this chapter (section 6.2), I assessed the robustness of the results from CCA, MMRM, LOCF and MI to variations in the trajectory profile (trajectory 1) and the magnitude of the treatment effect (−9.0 at the primary endpoint) considered in chapter 5 under various dropout scenarios. For CCA, MMRM and MI, the results were very similar across the trajectories and magnitudes of treatment effect when other factors were kept constant. However, the performance of LOCF varied considerably across the trajectories, especially when dropout rate was higher in the experimental group, where the improvement was substantial. It appears that the estimation of treatment effect using the LOCF approach was substantially affected by differential improvement over time between intervention groups. Hence, evaluation of trajectory profile by dropout patterns may be helpful to assess the impact of the LOCF approach.

In the second part of the chapter (section 6.3), I compared the results from MMRM with baseline-as-covariate to an alternative repeated measures analysis model with baseline-as-outcome. Since the presence of early dropouts is common in pragmatic trials, all simulated datasets in this thesis involved some participants without any follow-up measurements. Hence, analysis using the model with baseline-as-covariate led to the exclusion of those participants from the analysis. It was found that the inclusion of participants without any follow-up data (by considering baseline as an outcome) did not make a difference to the

estimates of treatment effect. However, with 30% dropouts, the study found a difference in the coverage of the 95% CI and the observed power. The model with baseline-as-covariate and Kenward-Roger correction showed the highest coverage, and the model with baseline-as-outcome and without the correction showed the lowest coverage. In contrast, the model with baseline-as-outcome and without Kenward-Roger correction showed the highest power, and the model with baseline-as-covariate and the correction showed the lowest power. The differences were noticeable for the dropout mechanism with differing directions compared with the same direction.

In the final part of this chapter (section 6.4), I evaluated the common practice of inflating sample size – by the inverse of one minus the anticipated dropout rate – to retain the desired power at a nominal level of 80% or 90% in the presence of dropouts. Overall, the inflation in sample size was helpful in protecting against the loss of power due to attrition. When the dropout rate was 10%, the methods – CCA, MMRM and MI – could retain the observed power very close to the desired level with the increased sample size in dropout scenarios where the estimate of treatment effect was unbiased. This was the case even with 30% dropouts under MCAR. However, the difference between the observed and desired power was noticeable in a few scenarios under MAR, though the estimate of treatment effect was unbiased. Importantly, the difference was not noticeable in the same-direction MAR mechanisms (MAR-B1 and MAR-L1), and was not substantially lower than 85% (when the desired power was 90%) or 75% (when the desired power was 80%) in cases where dropouts were in different directions between treatment groups (MAR-B2 and MAR-L2). Among the scenarios where the estimate of treatment effect was unbiased, the lowest statistical power was observed under MAR-B2.

6.6 Overall summary of findings from simulation studies

Simulation studies in chapters 5 and 6, which examined the relative performance of four statistical analysis approaches (CCA, LOCF ANCOVA, MMRM and MI ANCOVA), have demonstrated the impact of missing data on the estimation of treatment effect in an RCT. The comparison was in respect of various levels of experimental conditions typical of an RCT: levels of data variability, levels of correlation between repeated assessments, trajectory pattern, levels of overall dropouts, levels of differential dropout rates between groups, and direction of dropouts. These evaluations were performed in missing data scenarios where the missing data mechanism was known. Results from these simulation studies have shown that MMRM and MI yield unbiased estimates of treatment effect under MCAR and MAR-dependent on baseline or the last observed value, whereas CCA produces biased estimates with MAR-dependent on the last observed value. The performance of LOCF was severely influenced by the dropouts' trajectory profile and the timing of dropouts, irrespective of the missing data mechanism that was used to generate dropouts. That is, MMRM and MI ANCOVA are found to be more robust to bias from missing data compared to CCA and LOCF ANCOVA. Under all MNAR data scenarios, except the scenario of equal dropout rates with the same direction of dropout, none of the considered approaches performs well in terms of controlling bias in estimation of treatment effect; the bias markedly increases in relation to an increase in overall dropout rate and data variability. When the estimate of treatment effect from a statistical method was unbiased, the simulation studies showed that the coverage of 95% CI of the estimate was not affected (i.e. performance in respect of coverage closely aligned to performance in respect of bias); however, statistical power was substantially reduced with higher dropout. In the additional simulation studies, it was found that the inflation in sample size in proportion to the amount of dropout was helpful in protecting against the loss of power due to attrition. Also, the inclusion of participants without any follow-

up data (by considering baseline as an outcome with MMRM analysis) did not make a difference to the estimates of treatment effect compared to those from the analysis excluding such participants. However, the study found a slight difference in the coverage of the 95% CI and the observed power due to slightly reduced average SEs with the baseline-as-outcome method for scenarios of opposite direction of dropout.

Chapter 7: An empirical evaluation of the impact of missing data on treatment effect: analysis of TATE and STarT Back trials

7.1 Introduction

This chapter presents a re-analysis of two pragmatic clinical trials wherein the trial team had taken extra effort to minimize the amount of dropouts by sending reminders to initial non-responders and finally limiting data collection to key outcome measures. The present work utilizes the additional information from these trials and the earlier simulation study findings to assess the impact of missing data on the estimation of treatment effect. The next few sections (7.2–7.4) present the background, methodology and assumptions behind the present approach. Sections 7.5 and 7.6 present the results from the analysis of the two empirical datasets. Section 7.7 discusses the proposed approach and findings, and section 7.8 concludes the findings.

7.2 Background

Simulation studies in the previous chapters have shown that MMRM and MI yield unbiased estimates of treatment effect under MCAR and MAR dependent on baseline or the last observed value, whereas CCA produces biased estimates with MAR dependent on the last observed value. The performance of LOCF was severely influenced by the dropouts' trajectory profile and the timing of dropouts, irrespective of the missing data mechanism that was used to generate dropouts. In summary, the efficiency and accuracy of estimates from statistical methods depend on how close the mechanisms generating either the data or missing values are to the underlying statistical assumptions of the methods used. In practice, the missing data mechanisms are not strictly identifiable from incomplete data, so the desired "fit" in terms of assessing whether a method properly aligns to the mechanism of missingness is difficult to ascertain. Further, comparison of estimates of treatment effect from different methods in an empirical dataset is not

sufficient to make a valid conclusion about the unbiasedness of the estimates since the 'true' value is unknown. As seen in the simulation studies under MNAR mechanisms, the equality of estimates from these methods may not guarantee accuracy.

An MMRM or MI-based analysis should be treated as the primary analysis in longitudinal RCTs because of the following: (i) MMRM/MI-based analysis provides accurate and consistent estimates of treatment effect in relatively larger number of scenarios than CCA/LOCF-based analysis and (ii) since the observed outcome is likely to be associated with dropouts in a longitudinal RCT, an MAR missing data mechanism might be more plausible than an MCAR mechanism. Since it is not possible to guarantee that at least some of the missing data are not MNAR, it is important to assess the sensitivity of results from an MAR-based analysis to departure from the MAR assumption. The NRC report on the prevention and treatment of missing data in clinical trials (National Research Council, 2010) highlights the need for sensitivity analyses to confirm the primary analysis findings; however, the report also acknowledges the lack of guidelines on the selection of sensitivity analyses and interpretation of their findings, and lack of software packages[16] to implement such analyses. Considering the difficulties associated with the sensitivity analyses, I propose an approach using the responses obtained after a number of failed attempts to verify the ignorability of the missing data that is assumed by the primary analysis and hence the unbiasedness of the estimate of treatment effect.

7.3 Reminder responses as proxies of non-responses

In an effort to minimize the amount of dropout in pragmatic RCTs, trialists often follow a strategy of sending reminders to initial non-responders and re-approaching them for minimum data collection (MDC), where data collection is usually limited to the primary

[16] PROC MI in SAS/STAT(R) 13.2 (SAS Institute Inc., 2014) has a few options for doing a sensitivity analysis based on MI.

outcome variable plus perhaps other key outcome measures, if they have still failed to respond. The data that are recovered through the reminder strategy would otherwise have been missing. That is, one of the key features of the reminder responses is that they are data that can be treated as missing while knowing their true value. The simulation studies in chapters 5 and 6 showed that the estimate of treatment effect from a statistical method is not affected by the amount of dropout when the mechanism behind the missing data meets the assumption associated with the statistical method. For example, in MAR data, MMRM analysis yielded similar estimates of treatment effect irrespective of a 10% or 30% dropout rate; whereas in MNAR data, MMRM analysis yielded dissimilar estimates of treatment effect under 10% and 30% dropout rate scenarios. The purpose of this chapter is to explore empirically the mechanism behind the reminder responses in specific trial data by utilizing the key feature of the reminder data and the simulation findings, and thereby to verify the missing data mechanism that is assumed by the primary analysis. The following paragraphs explain the approach in detail.

The present approach considers two data scenarios: (i) one with the *actual* dataset and (ii) another one with a *modified* dataset, where outcome responses after a certain number of reminders are regarded as missing. The comparison of estimates from MAR-based analyses between the two data scenarios – actual versus modified datasets – identifies the impact of the reminder responses. If the mechanism behind the reminder responses meets the statistical assumption with regard to the statistical method applied for the estimation of treatment effect, the estimates of treatment effect from the actual and the modified datasets are expected to be similar. This conclusion is justified by the simulation studies (Chapters 5 and 6) in which it was found that a valid MAR-based estimate of treatment effect was not influenced by the amount of missing data that were generated under an MCAR or Mar mechanisms. Therefore, similar MAR-based estimates from the actual and modified datasets indicate that the mechanism behind the reminder responses is potentially ignorable. On the other hand, dissimilar estimates of treatment effect from

the actual and the modified datasets based on an MMRM or MI analysis indicate that either reminder or actual missing responses are non-ignorable under the MAR-based analysis. That is, dissimilar estimates from the datasets do not confirm a particular mechanism to either the reminder or actual missing responses.

In order to extend the finding on reminder responses to the actual missing responses, it is required to assume the reminder responses as representative of the actual missing responses. The plausibility of the assumption can be increased by defining the reminder responses as the responses that are obtained after a number of failed attempts to recover the data. If reminder responses are representative of the actual missing responses, then the mechanisms generating the reminder responses and the missing responses might be similar. Therefore, similar estimates of treatment effect based on an MAR-based analysis from the actual and the modified datasets generally indicate an ignorable missing data mechanism and the unbiasedness of the estimates obtained from both the datasets, if the assumption holds. Correspondingly, dissimilar estimates of treatment effect based on an MAR-based analysis from the actual and modified datasets generally indicate a non-ignorable missing data mechanism, if the assumption holds. In the latter case, the estimates of treatment effect from MAR-based analyses of the actual and modified datasets are biased, and the difference in estimates between the datasets might be the additional bias associated with the reminder responses in the modified dataset where the reminder responses are regarded as ‘missing’ for the comparison.

7.4 Methods

This study used datasets from two recent RCTs: (i) an RCT that evaluated transcutaneous electrical nerve stimulation as an adjunct to primary care management for tennis elbow (TATE trial; Chesterton et al., 2009; 2013) and (ii) an RCT that compared stratified primary care management for low back pain with current best practice (STarT Back trial; Hay et al., 2008; Hill et al., 2011). Both primary and key secondary outcome variables

from these trials were included since these variables provide different levels of missing data (depending on whether or not the MDC strategy had been employed) and can add to the imputation modelling of the MI method. In each trial dataset, initial explorations were carried out with respect to the amount and pattern of missing data. Forward logistic regression models were used to investigate which factors – such as baseline covariates, post-randomization variables and previously observed outcomes – were associated with the probability of missingness for outcomes at each follow-up. A significant finding counters the possibility of an MCAR mechanism. Further, correlation analysis was used to assess association between variables in each datasets. The predictors of missingness for an outcome and predictors of the outcome were used for evaluating an MI modelling strategy.

The incomplete empirical datasets were analysed using the four different approaches to deal with missing values. In the first approach, a standard ANCOVA (i.e. CCA) was employed, wherein a substantial number of participants who failed to provide sufficient data were excluded. In order to take account of the exclusion of participants with missing values from the ANCOVA model, incomplete datasets were also analysed using the approaches of: LOCF ANCOVA, MI ANCOVA and MMRM. MI was implemented in two different ways by considering two modelling strategies: *restrictive* and *inclusive*. In the *restrictive* imputation model, the imputation phase included only variables considered in the subsequent analysis (estimation) model; whereas the *inclusive* imputation model included auxiliary variables that were not part of the subsequent analysis model. The simulation studies in this thesis were very restrictive in that they utilized an MI imputation model that included only the outcome variable of interest and the treatment indicator (so being similar in content to the comparison MMRM model). Both the MMRM and MI-restrictive imputation analyses yielded similar estimates of treatment effect but slightly different SEs. Sequential imputation using chained equations (StataCorp, 2013) with a regression procedure was used to impute missing values in the outcome variables. Five-

hundred imputed datasets were created to reduce sampling variability from the imputation process and 50 iterations were used for the burn-in period prior to saving each imputed dataset to ensure a chain that converged to a stationary distribution. The MMRM with baseline-as-covariate analysis was performed using a restricted maximum likelihood estimation procedure. The model included the fixed-effect outcome variables, categorical effects of treatment and treatment-by-time interaction, as well as the fixed covariates of baseline score, age and sex, and covariates-by-time interactions. An unstructured covariance structure was used to model within-subject errors. Since the presence of early dropouts was substantial in both the trials, the MMRM analysis was repeated by considering baseline-as-outcome to make the analysis truly intention-to-treat and to explore the impact of the exclusion of participants with only baseline data.

The estimates of treatment effect at the final visit (primary endpoint) and its SE for the primary and the secondary outcome variables were obtained from these analysis methods in each data scenario (i.e., the actual and the modified datasets). The estimates were adjusted for age, sex and baseline score. Standardized effect size of treatment effect for an outcome was calculated as the estimate of treatment effect divided by the baseline SD of the outcome variable. The advantage of using baseline SD as the denominator is that the baseline spread is not influenced by dropouts since baseline observations are usually available on all randomized subjects.

The comparison of effect size from CCA with that from MMRM/MI analysis helps to assess the impact of the MCAR assumption with respect to the MAR assumption. Dissimilar estimates from methods assuming MCAR and MAR mechanisms can be taken as an indication against the MCAR mechanism; however, agreement between CCA and MMRM/MI analyses cannot be taken as an indication that data are MCAR. As found in the simulation studies, it is quite possible for a mechanism to produce non-ignorable missingness, yet result in comparable outcomes in these methods.

Finally a comparison of estimates from the actual and modified data scenarios was performed. Importantly, if the assumption that the reminder responses are representative of the actual missing responses holds, the difference in estimates between the two data scenarios can inform a preliminary estimate of the potential non-response bias associated with an MAR-based analysis of the actual dataset. In order to make the assumption more plausible, the reminder responses on the primary outcome were defined as the responses obtained through the MDC (i.e. responses retrieved after three failed attempts) in the TATE and STarT Back trials. The reminder responses on secondary outcome variables, where no MDC strategy was implemented, were defined as the responses obtained after the second reminder in the TATE trial; however, no reminder information on secondary outcome variables was available in the STarT Back trial. If the assumption holds, similarity in estimates between the data scenarios indicates an MCAR or MAR mechanism and unbiasedness of the estimates obtained from the actual data, whereas dissimilarity indicates an MNAR mechanism and the possibility of bias in the estimate of treatment effect from the actual data.

7.5 The TATE trial

The TATE trial (Chesterton et al., 2009; 2013) was designed as a pragmatic RCT to investigate the effectiveness of transcutaneous electrical nerve stimulation (TENS) as an adjunct to primary care management (PCM) for reducing pain intensity in patients with tennis elbow. Two hundred and forty-one participants with a first or new clinical diagnosis of tennis elbow were randomly allocated to either PCM alone (n = 120) or PCM plus TENS (n = 121). The total sample size was calculated to detect a 20% difference in the primary outcome between intervention groups with 90% power, 5% two-tailed significance level and 15% anticipated loss to follow-up. The primary outcome was the intensity of elbow pain, and the secondary outcomes were patient-rated tennis elbow evaluation (PRTEE) and the 12-item short-form health survey (SF12). The pain intensity score ranges from 0

(no pain) to 10 (worst pain imaginable). The PRTEE score ranges between 0–100, with high scores indicating greater pain/limitation. The SF12 provides two summary measures – physical component summary (SF12-PCS) and mental component summary (SF12-MCS), each measured on a 0–100 scale with high scores indicating better general health. All the outcomes were self-reported. The outcomes were assessed at baseline prior to randomization and were further measured at six weeks, six months, and 12 months after randomization by postal questionnaires. Though 6 weeks evaluation was the original primary endpoint, the 12 months is regarded as the primary endpoint in this analysis in order to ensure two interim visits prior to the primary endpoint.

At each follow-up if a participant did not return their questionnaire within two weeks after the first mail-out a reminder was issued, and subsequently a second reminder two weeks later if they had still not responded. Participants who did not respond to the reminder letters were telephoned twice: the first call as a reminder and the second one to those who did not respond to the previous reminders to collect minimum data (i.e. data on the primary outcome). The data that were collected through either the reminders or the MDC – the data that would otherwise have been missing – allows trialists to investigate the impact of the reminder strategy on the trial's conclusion. The average number of days delay in response after the first mail-out at each follow-up visit is provided in table 7.1. For the purpose of the re-analysis, participants who responded through the MDC strategy were considered as *'reminder'* responders on the primary outcome (pain intensity); the remaining responders were considered as *'initial'* responders for the primary outcome. For the secondary outcomes where there was no MDC, participants who responded to the first mail-out or the first reminder were considered as 'initial' responders, and participants who responded after the second mail reminder were considered as 'reminder' responders. This classification is based on the assumption that the reminder data are more likely to represent the non-responder data when the number of failed attempts (i.e. number of reminders) increases.

Table 7.1: Responders' status at follow-up assessments

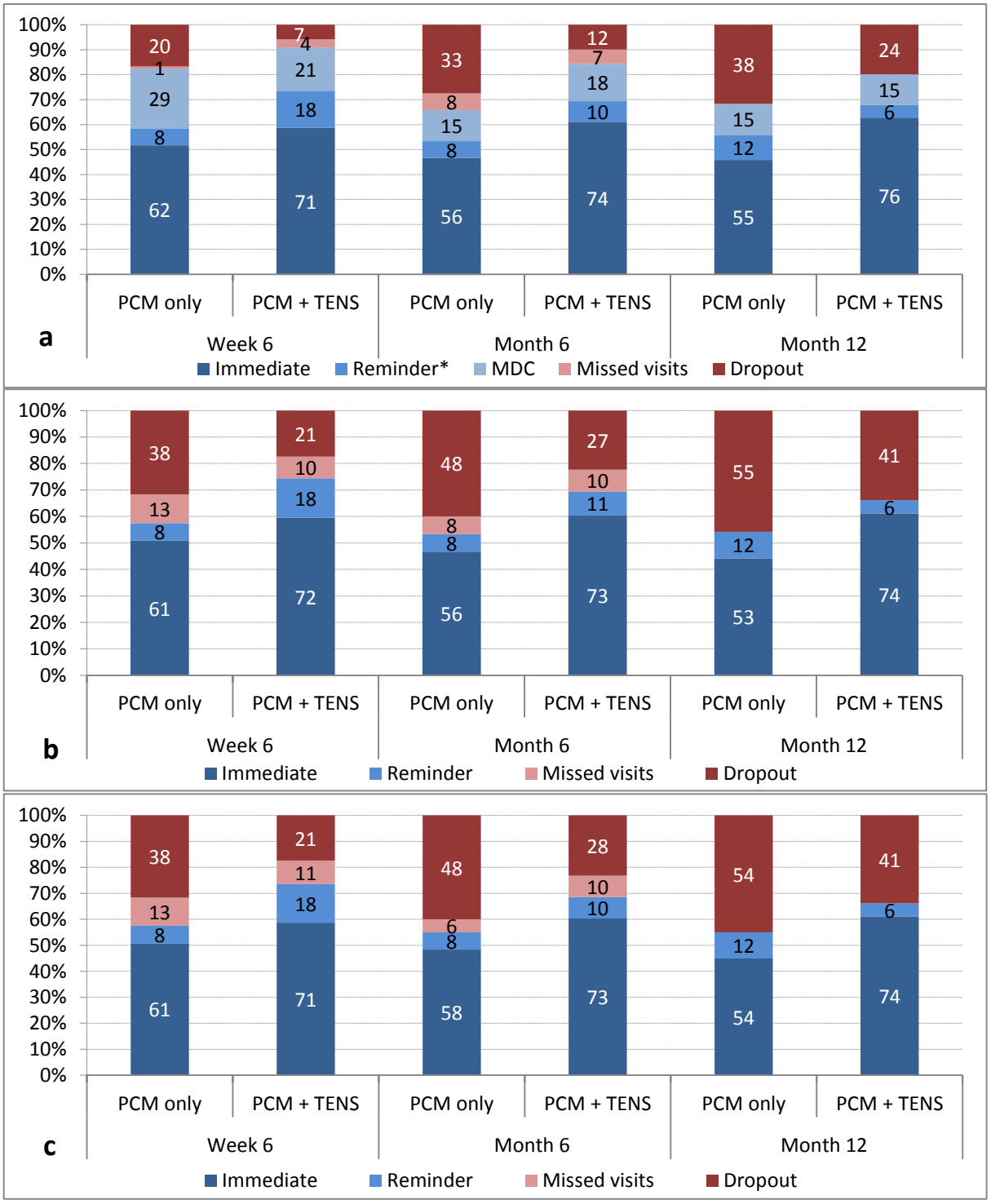
Responding through	Week 6		Month 6		Month 12	
	n	Mean*	n	Mean*	n	Mean*
First mail-out	93	11.5	109	14.1	103	13.2
First reminder	41	22.0	23	38.9	28	39.4
Second reminder†	-	-	14	56.0	13	55.8
First telephone call	26	56.4	4	114.8	5	76.8
Minimum data collection	50	90.9	31	110.0	26	92.7

* Average number of days delay in response after the first mail-out.
†There was no second reminder at week 6.

7.5.1 Descriptive analysis of missing data

7.5.1.1 Missing data in the TATE trial

The proportions of missing data for the primary and secondary outcome variables in the actual dataset are shown in figure 7.1. The dropout rate was up to 40% (96/241) at the final visit for some variables (Figure 7.1b) with differential rates between the two intervention groups (46% [55/120] in PCM only and 34% [41/121] in PCM plus TENS). A better response rate was achieved for the pain intensity score through the MDC strategy (Figure 7.1a). The dropout rate for the pain intensity score at the final visit was 26% (32% [38/120] in PCM only and 20% [24/121] in PCM plus TENS). As seen in the figures, a substantial number of participants dropped out even prior to the first follow-up at week 6, and intermittent missing data (due to missed visits or missed item) were limited to a small proportion of participants in all outcome variables. The trial could obtain reasons for dropouts in limited instances only. The reasons could have been helpful in guiding judgement on the missing data mechanism if it were obtained in full.



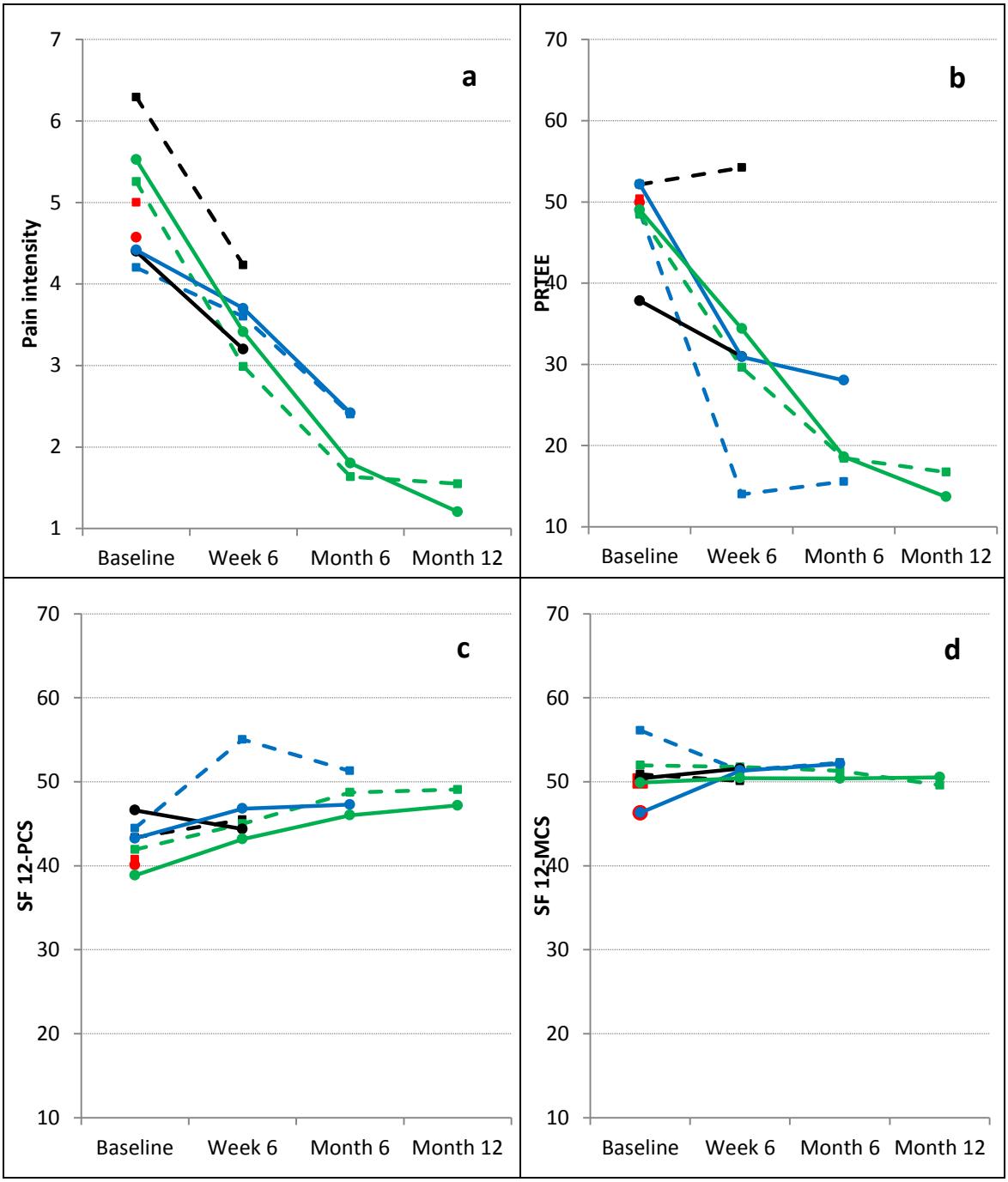
The numbers displayed on the bars represents the number of participants; the reminder data includes the responses recovered after second reminder; *reminder data excludes the responses through MDC

Figure 7.1: Response rate (%) over time on outcome measures – (a) pain intensity, (b) PRTEE total score, and (c) SF12 score.

In the modified dataset, the amount of missing data at the final visit for the primary outcome, where minimally collected responses (retrieved through MDC strategy) were designated as missing, was 44% (53/120) in PCM only and 32% (39/121) in PCM plus TENS. That is, a difference of 12% in dropout rate between the actual (26%; 62/241) and the modified (38%; 92/241) datasets. For the secondary outcome measures where reminder responses (retrieved through second and third reminders) were designated as missing, it was 56% (67/120) in PCM only and 39% (47/121) in PCM plus TENS. That is, a difference of 7% in dropout rate between the actual (40%; 96/241) and modified (47%; 114/241) datasets.

7.5.1.2 Dropout pattern and predictors of missingness in the TATE trial

Knowing the missing data mechanism is important in determining the best way to handle the missing data in order to provide the least biased results. Figure 7.2 displays the mean score over time by different dropout patterns for the primary and secondary outcome variables. The number of participants within each dropout pattern can be determined from figure 7.1. Taking figure 7.2a as an example, 20 and 7 participants dropped out before the first follow-up assessment from the control and intervention groups, respectively. The observed mean baseline scores for these participants are represented with square and round dots in figure 7.2a. Another 18 participants (13 from the control group and 5 from the experimental group) were lost to follow-up in between week 6 and month 6. The group-wise mean score at baseline and week 6 for these participants are represented in lines (black lines; solid for PCM plus TENS and dashed for PCM only) ending at week 6. Another 17 (5 from the control group and 12 from the experimental group) dropped out before the endpoint at month 12. The group-wise mean scores at baseline, week 6 and month 6 for these participants are represented in lines (blue lines) ending at month 6. The green lines represent the completers.



The solid lines represent the estimates from the PCM plus TENS group and dashed lines represent the estimates from PCM only group. Red denotes dropouts between baseline and week 6; black denotes dropouts between week 6 and month 6; blue denotes dropouts between month 6 and month 12; green denotes completers.

Figure 7.2: Observed mean profiles according to intervention groups and time at which participants lost to follow-up.

If mean profiles are similar across different dropout patterns, then the data help support the MCAR missingness mechanism; otherwise, if the mean trajectories are dissimilar, this would counter MCAR in favour of MAR or MNAR. For the pain intensity score (Figure 7.2a), the mean profiles for the dropouts at various time-points were notably different from those who completed the final follow-up. Further, the observed differences between the intervention groups in the last observed values before dropping out were dissimilar across the dropout pattern. The differential dropout rate, combined with the trend of improved pain intensity score for all subjects in later visits (i.e. over time), would cause LOCF analysis to give inaccurate estimates of the treatment effect at the final visit. A similar observation could be made with other variables: PRTEE (Figure 7.2b) and SF12-PCS (Figure 7.2c). Importantly, in figure 7.2b, the dissimilarity in the observed differences across the dropout patterns was noticeable, and hence the estimates of treatment effect from CCA and LOCF approaches would be expected to be different. In figure 7.2d, the dissimilarity was not noticeable across the dropout patterns except for those who dropped out without any follow-up measurements. Due to a significant number of early dropouts, the difference in the mean baseline value of SF12-MCS for the early dropouts between the intervention groups (50.1 for PCM only and 46.3 for PCM plus TENS) may influence the LOCF estimate of treatment effect.

The odds ratio (OR) from logistic regression models indicated that the intervention group was significantly associated with a lower probability for missing response in the primary outcome (adjusted OR = 0.44, p = 0.039), where MDC strategy had been applied, at the first follow-up (week 6). For the secondary outcomes, where the level of missing data was similar across the outcomes and the MDC strategy had not been applied, the intervention group (adjusted OR = 0.37, p = 0.001), and increased baseline age (adjusted OR = 0.95, p = 0.002) and SF12-MCS (adjusted OR = 0.97, p = 0.049) showed an association with the lower level of missingness.

At month 6, the intervention group and age at baseline were significant predictors of missingness in the primary outcome and the secondary outcomes. The adjusted OR for the intervention group was 0.30 (p-value < 0.001) for missingness in the primary outcome, and 0.39 (p-value = 0.001) for missingness in the secondary outcomes; for age at baseline, the adjusted OR was 0.96 (p = 0.025) and 0.93 (p < 0.001), respectively. The intervention group (adjusted OR = 0.10, p = 0.001), baseline pain intensity (adjusted OR = 1.82, p = 0.017), and week 6 PRTEE (adjusted OR = 1.08, p = 0.006) and SF12-PCS (adjusted OR = 1.15, p = 0.033) were found to be significantly associated with the additional missing responses at month 6 compared to week 6. The probability of the additional missing responses in the secondary outcomes was dependent on baseline data on age (adjusted OR = 0.94, p = 0.025), pain intensity (adjusted OR = 1.42, p = 0.043) and PRTEE (adjusted OR = 0.93, p = 0.005) and week 6 data on PRTEE (adjusted OR = 1.03, p = 0.030).

The intervention group and age were significant predictors of missing response in the primary and secondary outcomes at month 12. However, no variables (either baseline or previously observed data) were found to be significantly associated with the additional missing responses at month 12 compared to month 6.

In summary, the above associations provide evidence against the MCAR scenario in the dataset; however, there is no evidence to favour either an MAR or an MNAR mechanism.

Table 7.2: The observed pairwise correlation between variables in the actual dataset

		Age	Pain score				PRTEE				SF12-PCS				SF12-MCS			
			w0	w6	m6	m12	w0	w6	m6	m12	w0	w6	m6	m12	w0	w6	m6	m12
Age		1.00																
Pain score	w0	0.03	1.00															
	w6	-0.07	**0.40**	1.00														
	m6	-0.06	0.16	**0.37**	1.00													
	m12	-0.12	0.24	**0.33**	**0.58**	1.00												
PRTEE	w0	0.01	**0.72**	**0.34**	0.21	0.26	1.00											
	w6	-0.13	**0.47**	**0.85**	**0.37**	**0.35**	**0.53**	1.00										
	m6	-0.03	0.23	**0.53**	**0.85**	**0.61**	**0.35**	**0.60**	1.00									
	m12	-0.09	0.27	**0.40**	**0.65**	**0.92**	**0.34**	**0.48**	**0.73**	1.00								
SF12-PCS	w0	-0.17	**-0.40**	-0.17	-0.12	-0.20	**-0.46**	**-0.35**	-0.26	**-0.32**	1.00							
	w6	-0.18	-0.24	**-0.36**	-0.22	-0.14	**-0.34**	**-0.45**	**-0.38**	-0.27	**0.62**	1.00						
	m6	-0.10	-0.08	-0.27	**-0.40**	-0.23	-0.22	-0.29	**-0.49**	**-0.36**	**0.46**	**0.67**	1.00					
	m12	-0.08	-0.23	-0.22	**-0.36**	**-0.32**	**-0.38**	-0.29	**-0.47**	**-0.46**	**0.59**	**0.64**	**0.75**	1.00				
SF12-MCS	w0	0.14	-0.14	-0.17	-0.25	-0.16	-0.23	-0.18	-0.29	-0.14	0.00	0.10	0.26	0.26	1.00			
	w6	0.17	-0.08	-0.11	-0.07	0.02	-0.22	-0.29	-0.18	-0.03	0.16	0.08	0.18	0.18	**0.63**	1.00		
	m6	0.06	-0.10	-0.10	-0.10	-0.09	-0.14	-0.14	-0.18	-0.09	0.13	0.13	0.04	0.13	**0.50**	**0.59**	1.00	
	m12	0.05	-0.14	-0.03	-0.04	-0.15	-0.16	-0.17	-0.16	-0.20	0.22	0.09	0.06	0.05	**0.56**	**0.55**	**0.61**	1.00

Bold values represent the absolute values of the correlations higher than 0.30.

Table 7.2 provides the observed pairwise correlation between variables at different occasions. Though age was a significant predictor of missingness in outcome variables, it did not show a significant correlation with any of the outcome variables at any occasion. PRTEE showed the strongest correlation with pain intensity and SF12-PCS; however, SF12-MCS showed only weak correlations with any other variables.

In the next section, both the actual and the modified datasets were analysed using the considered statistical methods, and an assessment made of the impact of the reminder response on the trial's conclusion from the data with no distinction between the initial and reminder data.

7.5.2 Analysis of the incomplete TATE trial – estimation of treatment effect at month 12

7.5.2.1 Results from the actual dataset

Table 7.3 provides the estimates of treatment effect, SE and standardized effect size for the primary and secondary outcomes at the final visit based on a standard ANCOVA (i.e. CCA) and LOCF ANCOVA models. Due to an MDC strategy, the number of participants with complete data on the primary outcome (pain intensity) was substantially higher compared to the secondary outcome measures. As shown in table 7.3, LOCF produced a substantially different estimate of treatment effect in comparison with CCA, particularly for the pain and function scales.

Table 7.3: TATE - ANCOVA results at 12 months follow-up before and after LOCF imputation of missing values

Outcome variables[1]	Standard ANCOVA[2]				LOCF ANCOVA[3]			
	Estimate	SE	p-value	Standardized effect size[4]	Estimate	SE	p-value	Standardized effect size[4]
Pain intensity	-0.452	0.302	0.137	-0.220	-0.953	0.304	0.002	-0.464
PRTEE	-3.599	3.057	0.241	-0.203	-8.566	2.841	0.003	-0.483
SF12-PCS	0.134	1.476	0.928	0.014	1.073	1.089	0.326	0.111
SF12-MCS	2.214	1.532	0.151	0.202	1.900	1.057	0.073	0.173

Estimate – estimate of treatment effect for an outcome at month 12 adjusted for age, sex, baseline pain intensity, and the corresponding baseline of the outcome; SE – standard error; [1]Pain intensity measured on a 0–10 scale, other outcomes on a 0–100 scale. [2]Number of subjects included in the analysis was 179 for pain intensity, 145 for PRTEE, and 146 for SF12-PCS and MCS. [3]Number of subjects included in the analysis was 241 for all outcomes. [4]Treatment effect relative to the pooled SD of baseline scores.

Under the standard ANCOVA analysis, the estimates of treatment effect in both the primary and secondary outcomes were statistically non-significant at the primary endpoint. However, the estimates of treatment effect of pain intensity (measured on a 0–10 scale) and PRTEE (measured on a 0–100 scale) markedly favoured the new intervention when missing data were imputed through LOCF approach – the absolute difference in standardized effect size between the two analysis methods was 0.244 for pain intensity and 0.280 for PRTEE measure. The treatment effect in these two outcome measures became statistically significant with LOCF. The estimates of treatment effect in SF12-PCS (measured on a 0–100 scale) also increased in favour of the new intervention at the primary endpoint under LOCF ANCOVA but retained non-significance. On the other hand, the estimate for SF12-MCS (measured on a 0–100 scale) slightly reduced under LOCF ANCOVA compared to the standard ANCOVA (though the p-value was also slightly lower). LOCF ANCOVA yielded a lower SE compared to the standard ANCOVA in all scenarios, except for pain intensity at month 12, in relation to the amount of missing data.

Table 7.4 presents the treatment effect in the primary and the secondary outcomes at the final visit estimated from the actual dataset based on MMRM models. The estimate of

treatment effect did not differ by the baseline handling strategies in the MMRM model. That is, subjects with only baseline values did not make a difference to the estimate at the follow-up visit. However, the SE of the estimate was slightly lower in an MMRM model with baseline-as-outcome compared to the baseline-as-covariate strategy.

Table 7.4: TATE - MMRM results

Outcome variables[1]	MMRM (baseline-as-covariate)[2]				MMRM (baseline-as-outcome)[3]			
	Estimate	SE	p-value	Standardized effect size[4]	Estimate	SE	p-value	Standardized effect size[4]
Pain intensity	-0.456	0.300	0.129	-0.222	-0.456	0.299	0.127	-0.222
PRTEE	-3.788	2.910	0.193	-0.214	-3.788	2.900	0.191	-0.214
SF12-PCS	-0.193	1.422	0.892	-0.020	-0.193	1.396	0.890	-0.020
SF12-MCS	2.502	1.505	0.096	0.228	2.503	1.494	0.094	0.229

Estimate – estimate of treatment effect at month 12 adjusted for fixed covariates (age, sex, baseline pain intensity, baseline of outcome variable) and their interaction with time; SE – standard error; [1]Pain intensity measured on a 0–10 scale, other outcomes on a 0–100 scale. [2]Number of subjects included in the analysis was 214 for pain intensity and 182 for other outcomes. [3]Number of subjects included in the analysis was 241 for all outcomes. [4]Treatment effect relative to the pooled SD of baseline scores.

For the primary outcome (pain intensity), the estimate of treatment effect and its SE from MMRM analysis were very close to that from CCA. For the secondary outcomes (PRTEE, SF12-PCS and SF12-MCS), the standardized effect sizes were slightly more in favour of the new intervention compared to CCA. In combination with a reduced standard error, this resulted in slightly lower p-values for MMRM analysis compared to CCA, but this small change was not sufficient to influence statistical significance.

Table 7.5 presents the MI-based ANCOVA results from the actual dataset. A restrictive imputation model included only the variables considered for an MMRM model for an outcome variable (i.e. baseline score, age and sex in addition to the outcome variable), whereas an inclusive model additionally included variables that were not part of the MMRM model for that outcome variable. The estimates of treatment effect from the MI ANCOVA with restrictive modelling were comparable to that from MMRM for all outcome

variables. However, the SE of the estimate from the MI was higher than that from MMRM, but comparable to that from CCA.

Table 7.5: TATE - ANCOVA results after MI imputation of missing values

Outcome variables[1]	MI (restrictive modelling)[2]				MI (inclusive modelling)[2]			
	Estimate	SE	p-value	Standardized effect size[3]	Estimate	SE	p-value	Standardized effect size[3]
Pain intensity	-0.458	0.305	0.136	-0.223	-0.497	0.306	0.106	-0.242
PRTEE	-3.817	3.038	0.211	-0.215	-3.961	2.771	0.155	-0.223
SF12-PCS	-0.276	1.448	0.849	-0.029	-0.424	1.500	0.778	-0.044
SF12-MCS	2.411	1.541	0.120	0.220	2.796	1.584	0.080	0.255

Estimate – estimate of treatment effect at month 12 adjusted for age, sex, baseline pain intensity, and corresponding baseline of outcome variable; SE – standard error of the difference; Restrictive modelling – imputation models included only those variables considered for the MMRM analysis models; Inclusive modelling – imputation model for an outcome included, in addition to the variables considered for the MMRM analysis model, other outcome variables as auxiliary variables in order to improve the performance of the imputation procedure; [1]Pain intensity measured on a 0–10 scale, other outcomes on a 0–100 scale. [2]Number of subjects included in the analysis was 241 for all outcomes. [3]Treatment effect relative to the pooled SD of baseline scores.

An inclusive imputation model with MI produced an increased standardized effect size for all outcome measures compared to the restrictive modelling based MI. Reduced SE for PRTEE was observed with the inclusive imputation modelling; however, there was a very slight increase in SE with the other outcome measures. Even though none of the auxiliary variables in the inclusive imputation model were found to be a significant predictor of missingness in SF12-MCS at month 12 or correlated with SF12-MCS, there was a difference of 0.035 in the standardized effect size for SF12-MCS and a difference of 0.043 in the SE between the two imputation modelling strategies. Results based on various inclusive imputation models with MI are reported in appendix 8 (Tables 27–30). The results indicate slight inconsistency in the estimate of treatment effect and SE, though none that affected statistical decision-making.

7.5.2.2 Results from the modified dataset and the impact of reminder or MDC responses

Table 7.6 provides the results from the modified dataset including deviation in the standardized effect size and SE from that observed with the actual dataset. As observed with the actual dataset, the estimates of treatment effect (and hence the standardized effect size) obtained from MMRM and MI with restrictive imputation modelling did not differ for any outcome measure in the modified dataset, and the SE was slightly lower for MMRM compared to MI. However, this was not the case with CCA versus MMRM or MI – both the estimates and SEs differed between the methods. The SE of the estimate from all methods increased due to the higher number of missing responses with the modified dataset.

In MAR-based analyses (MMRM and MI), an increased standardized effect size – approximately 3% of baseline SD – for pain intensity was observed with the modified data compared to the actual data; CCA also yielded a similar increase – approximately 4% of baseline SD – with the modified data. Similarly, an increased standardized effect size – approximately 5% of baseline SD – for PRTEE was observed with MAR-based analyses of the modified data, but little difference with CCA. With MMRM analysis, the standardized effect size for SF12-PCS was changed from −0.020 in the actual data to 0.033 in the modified data (i.e. approximately 5% of baseline SD). MI also yielded a similar difference; but, little difference with CCA. The standardized effect size for SF12-MCS differed between the two datasets mostly with MMRM and CCA: the estimate was reduced with MMRM while it increased with CCA. Importantly, for any outcome measure in the modified dataset, the final statistical conclusion did not change from the inference that was made in the actual dataset (p-value > 0.05 for all outcome measures).

Table 7.6: Results from the modified dataset

Outcome variables[1]	Modified dataset[2]					Deviation in results from actual dataset	
	n	Estimate	SE	p-value	Standardized effect size[3]	SE	Standardized effect size
Results from MMRM (baseline-as-covariate)							
Pain intensity	241	-0.524	0.327	0.109	-0.255	-0.028	0.033
PRTEE	241	-4.722	2.975	0.112	-0.266	-0.065	0.053
SF12-PCS	241	0.316	1.576	0.841	0.033	-0.154	-0.053
SF12-MCS	241	2.172	1.586	0.171	0.198	-0.081	0.030
Results from MI (restrictive modelling)							
Pain intensity	241	-0.528	0.331	0.112	-0.215	-0.026	0.034
PRTEE	241	-4.723	2.987	0.116	-0.266	0.051	0.051
SF12-PCS	241	0.397	1.629	0.808	0.041	-0.182	-0.070
SF12-MCS	241	2.233	1.627	0.173	0.204	-0.087	0.016
Results from ANCOVA							
Pain intensity	149	-0.533	0.336	0.115	-0.260	-0.034	0.039
PRTEE	127	-3.558	3.162	0.263	-0.201	-0.105	-0.002
SF12-PCS	128	0.192	1.665	0.908	0.020	-0.189	-0.006
SF12-MCS	128	2.464	1.616	0.130	0.225	-0.084	-0.023

n – number of subjects included in the analysis; estimate – estimate of treatment effect at month 12; SE – standard error; [1]Pain intensity measured on a 0–10 scale, other outcomes on a 0–100 scale. [2]reminder responses were regarded as missing (for pain intensity, the reminder responses include responses retrieved through MDC; for other outcomes, reminder responses include responses retrieved after second reminder). [3]Treatment effect relative to the pooled SD of baseline scores.

7.5.3 Summary and interpretation of findings

The estimates of treatment effect and SE with LOCF ANCOVA for pain intensity, PRTEE and SF12-PCS were quite different from those of the other imputation methods and CCA in the actual dataset. Variance estimates from CCA, MMRM and MI are generally larger than LOCF, because LOCF often underestimates standard errors when there is missing data. This study has, however, found an example where the LOCF-estimated variance for

pain intensity was higher than CCA, MMRM, or MI when the LOCF values were very different from the observed values at the final visit.

In the actual dataset, both the estimate of treatment effect and the SE for the pain intensity were similar across the MCAR-based and MAR-based analyses. However, based on this finding alone one should not necessarily conclude that the missing data mechanism was MCAR. Further, findings from the graphical evaluation of dropout patterns and logistic regression based identification of predictors of missingess did not support the MCAR assumption. The comparison of estimates from the actual dataset (dropout rate at month 12 was 26%) and the modified dataset (dropout rate at month 12 was 38%) found that treatment effect in pain intensity would have been overestimated by 3% of baseline pain intensity SD with an MAR-based analysis if an MDC strategy had not been implemented. This finding points towards a potential MNAR mechanism associated with missing pain intensity scores if the responses that had been retrieved through the MDC represent the actual missing data. It implies that the 'true' estimate of treatment effect (and effect size) might be lower than the observed estimate if the assumption holds; however, it was very unlikely that difference in the estimates would influence the statistical decision-making due to the small observed effect size.

For the secondary outcome measures, the amount of missing data differed only by around 7% between the actual dataset (amount of missing data at month 12 was 40%) and the modified dataset (amount of missing data at month 12 was 47%). However, similar to the finding on the primary outcome, all the secondary outcome measures showed a difference between the two datasets under MMRM analysis. The results from the MMRM analysis of the modified dataset showed that the estimate of treatment effect was overestimated by 5% of baseline SD for PRTEE and SF12-PCS, and underestimated by 3% of baseline SD for SF12-MCS, compared to the estimates from the actual dataset. These findings might be an indication of potential MNAR mechanisms associated with missing data in those

outcome measures if the responses (7%) that had been retrieved through the second and third reminders represent the actual missing data (40%). However, the representativeness can be criticized because of the large gap in the percentages and a minimal number (one) of failed attempts before obtaining the response, compared to the situation in the primary outcome. In addition, it was found that the absence of responses that had been retrieved through second and third reminders and the absence of minimally collected data had contrasting effects on the actual observed estimate of treatment effect in the primary outcome; i.e. MMRM analysis overestimated the actual estimate in the absence of minimally collected data whereas the similar analysis underestimated the actual estimate in the absence of the reminder responses (results are provided in appendix 9 [Table 31]). Though the findings on the secondary outcome measures provide an indication of potential bias associated with the actual missing responses, it is difficult to make a judgement of the direction of bias if it is assumed that minimally collected data might be more representative of the actual missing responses. Thus, these results show that it is important to investigate the robustness of the findings from an MAR-based analysis (MMRM) of these outcome measures in the actual dataset to the possibility of an MNAR mechanism behind the missing responses.

The findings from the TATE trial were generally comparable to the findings from the simulation studies. The analysis of the trial dataset confirms that the inclusion of participants without any follow-up data in MMRM analysis (by considering baseline as an outcome) did not make a difference to the estimates of treatment effect at the endpoint but there was a small reduction in SE. As found in the simulation studies, MI ANCOVA by restrictive imputation modelling did not add any advantage to the MMRM analysis. The estimate of treatment effect from the MI ANCOVA was close to that from the MMRM analysis when the number of imputations increased substantially. In this data, I used a large number of imputations, which was more than ten-fold of the current recommendation (White et al., 2011b), and a large number of iterations in the burn-in period, which was

fivefold the default value in Stata (StataCorp, 2013), to obtain a closer value of the estimates between the two analysis methods. However, as also found in the simulation studies, the SE with MI ANCOVA was still higher than that from the MMRM analysis. Further, it was found that the inclusive imputation modelling with MI resulted in a slight difference in the estimates of treatment effect and SE compared to restrictive imputation modelling with MI or MMRM analysis. The observed difference varied by the number of auxiliary variables added into the inclusive imputation model even in situations where the auxiliary variable was not associated with an outcome variable.

7.6 The STarT Back trial

The STarT Back trial (Hay et al., 2008; Hill et al., 2011) was designed to compare the effectiveness of stratified primary care for low back pain with current best practice. Participants were randomized to receive either a screening and targeted intervention, delivered by trained physiotherapists (intervention group; n = 568), or best current care (control group; n = 283). The total sample size (n = 851) was calculated with adjustment for 25% loss to follow-up at the primary endpoint (month 12). The primary outcome was the Roland Morris Disability Questionnaire (RMDQ), and the secondary outcomes were back pain intensity and health-related quality of life (using SF12). The RMDQ score ranges between 0–24, with high scores indicating severe disability. Back pain intensity was measured on a 0–10 numerical rating scale, with high scores indicating severe pain. As previously noted, the SF12 provides two summary measures – physical component summary (SF12-PCS) and mental component summary (SF12-MCS), each measured on a 0–100 scale, with high scores indicating better general health. The outcomes were obtained before randomization, and 4 months and 12 months later by use of postal questionnaires. The 12-month evaluation was regarded as the primary endpoint in this analysis.

In order to increase response rate, a similar reminder strategy was implemented as in the TATE trial. Participants who did not respond to the reminders were telephoned for MDC on the primary outcome. For the purpose of the re-analysis, participants who responded through the MDC strategy were considered as 'reminder' responders on the primary outcome (RMDQ); the remaining responders were considered as 'initial' responders for the primary outcome. Since details on reminders were not accessible electronically, this analysis did not consider the secondary outcomes, where there was no MDC, to investigate the impact of the missing data on the study conclusion.

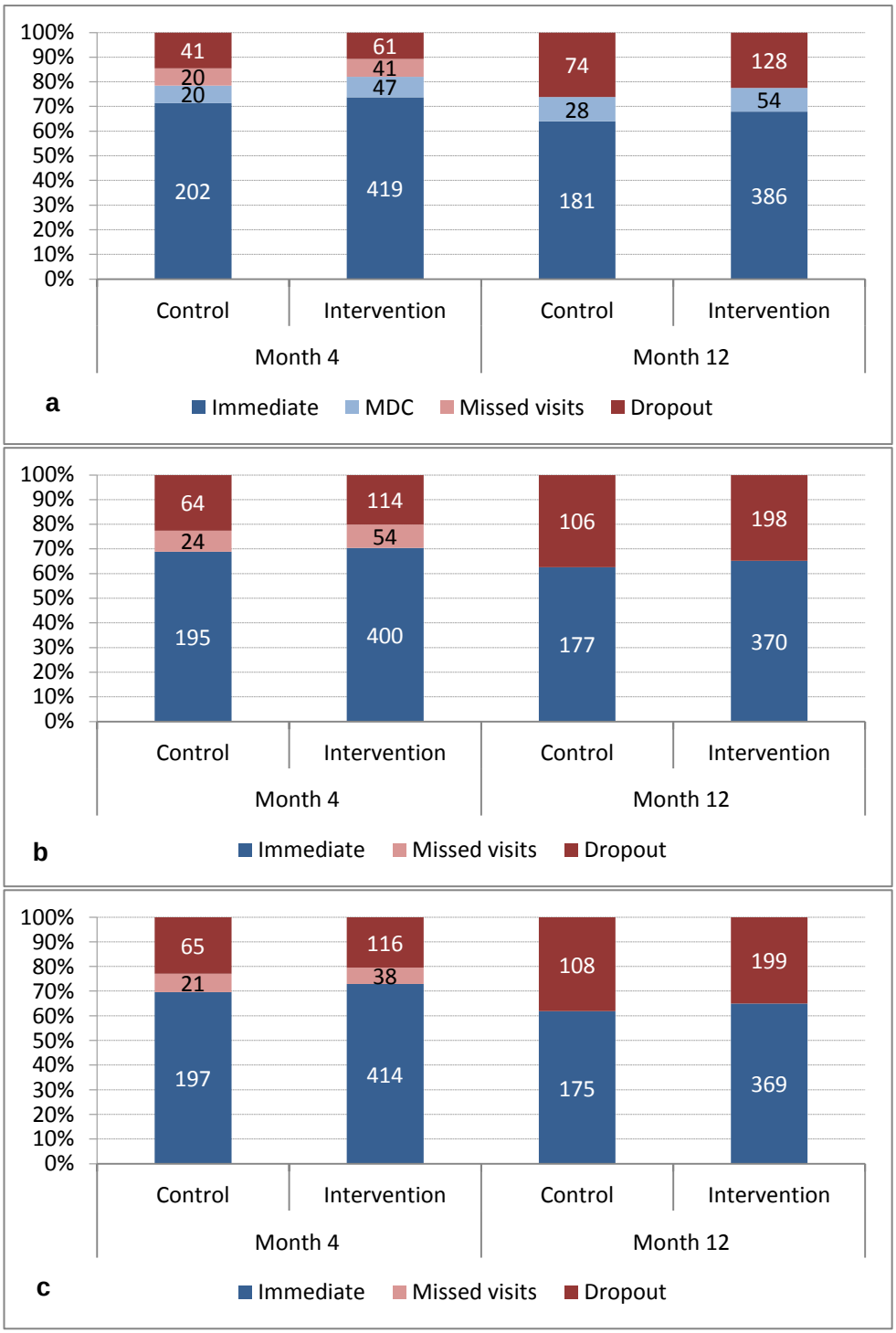
7.6.1 Descriptive analysis of missing data

7.6.1.1 Missing data in the STarT Back trial

The proportions of missing data for the primary and secondary outcome variables in the dataset are shown in figure 7.3. The dropout rate was up to 36% for the secondary outcome variables (Figure 7.3b [304/581] and figure 7.3c [307/851]) with nearly equal dropout rate between the two treatment groups (38% in the intervention group and 35% in the control group) at the final visit. A better response rate was achieved for the primary outcome (RMDQ) through the MDC strategy (Figure 7.3a): 10% (82/851) of randomized participants responded through this strategy at the final visit. The dropout rate for the RMDQ at the final visit was 24% (202/851; 26% in the intervention group and 23% in the control group). As seen in the figures, a substantial number of participants dropped out even before the first follow-up assessment at month 4. The trial team could obtain the reason for dropouts in limited instances only. Obtaining reasons would have been helpful in making a judgement on the missing data mechanism.

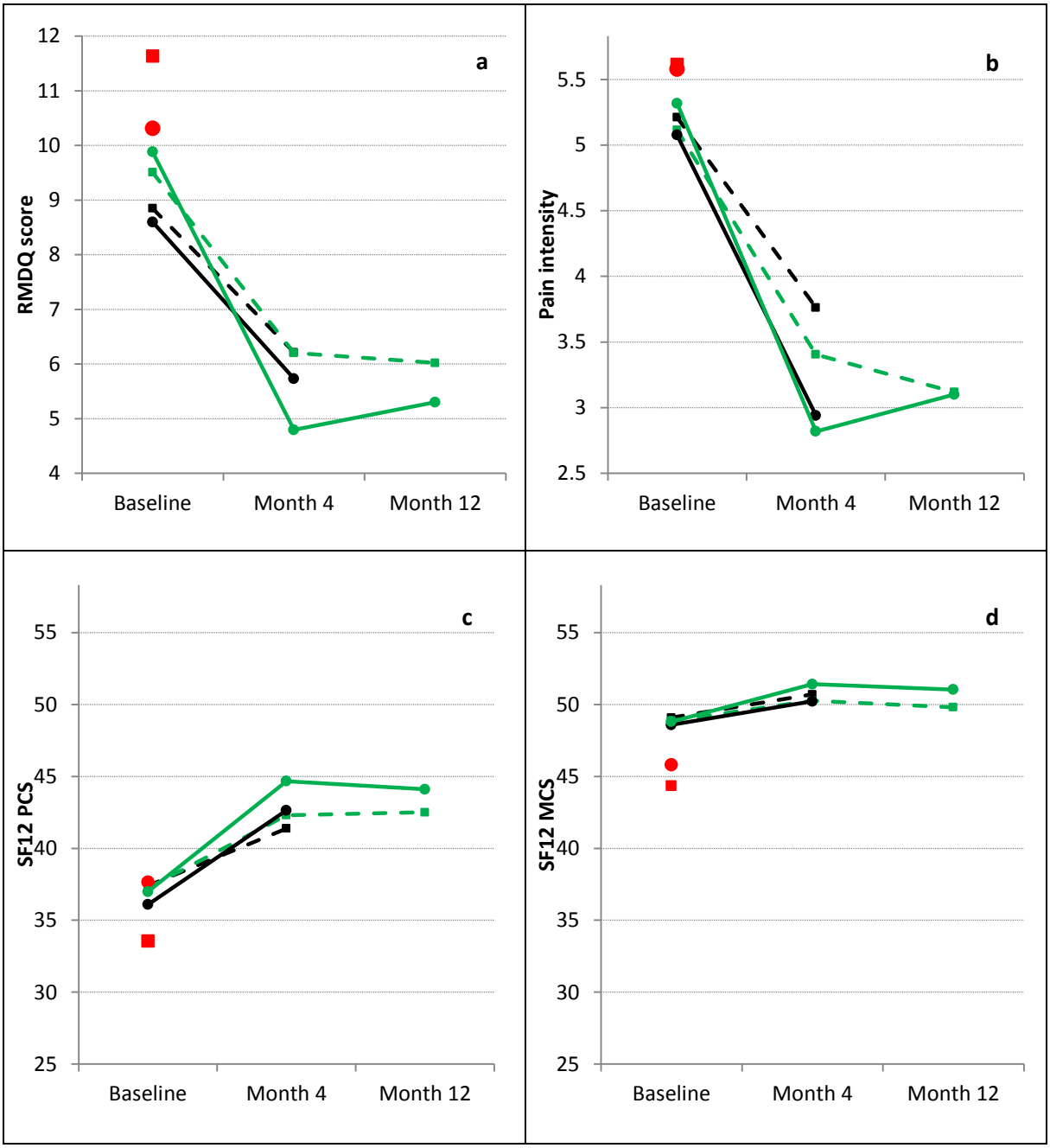
7.6.1.2 Dropout pattern and missing data mechanism in the STarT Back trial

Knowing the missing data mechanism is important in determining the best way to handle the missing data in order to provide the least biased results. Figure 7.4 displays the observed mean score over time by different dropout patterns for the primary and secondary outcome variables. The number of participants within each dropout pattern can be determined from figure 7.3. Taking figure 7.3a as an example, 41 and 61 participants dropped out before the first follow-up assessment from the control and intervention groups, respectively. The observed mean baseline scores for these participants are represented with square and round dots in figure 7.4a.



The number displayed on the bar indicates the number of responses

Figure 7.3: Response rate (%) over time on outcome variables – (a) RMDQ, (b) back pain intensity, and (c) SF12 (PCS & MCS)



The solid lines represent the estimates from the intervention group and dashed lines represent the estimates from the control group. Red denotes dropouts between baseline and month 4; black denotes dropouts between month 4 and month 12; green denotes completers

Figure 7.4: Observed mean profile according to intervention groups and the time at which dropped out

Another 100 participants (33 from the control group and 67 from the experimental group) dropped out after the first follow-up assessment; hence, no outcome measurements on these participants were obtained at the final follow-up. The group-wise mean score at baseline and month 4 for these participants were represented in lines ending at month 4.

Generally, those participants who dropped out immediately after baseline assessment displayed the worst score at baseline on all the outcome variables. In figure 7.4a, mean RMDQ score at month 4 in the intervention group appeared to be higher among those patients who did not provide data at month 12 than those who completed the assessment at month 12. Further, pain intensity score at month 4 in the control group appeared to differ between those patients who did and did not provide data at month 12 (Figure 7.4b).

Logistic regression models that were used to identify the possible observed predictors of missingness revealed that age (adjusted OR = 0.94, $p < 0.001$) and back pain intensity (adjusted OR = 1.12, $p = 0.050$) at baseline were significant predictors of missingness in the primary outcome, where MDC strategy had been employed, at month 4. For the secondary outcomes, where MDC strategy had not been employed, only age (adjusted OR = 0.95, $p < 0.001$) showed a significant association with missingness; the baseline back pain intensity measure was not a significant predictor for missingness (adjusted OR = 1.10, $p = 0.058$).

At month 12, age (adjusted OR = 0.95, $p < 0.001$) was the only significant predictor of the missing response in the primary outcome; whereas age (adjusted OR = 0.94, $p < 0.001$) and baseline SF12-PCS (adjusted OR = 0.98, $p = 0.026$) were significant predictors of the missing response in the secondary outcomes. When the missing response was restricted to dropouts after month 4, age (adjusted OR = 0.97, $p = 0.001$) was still associated with the additional dropouts with respect to the primary outcome. However, for the additional dropouts with respect to the secondary outcomes, SF12-MCS at month 4 (adjusted OR = 0.98, p-value = 0.028) was a predictor in addition to these baseline variables: age

(adjusted OR = 0.96, $p < 0.001$), RMDQ (adjusted OR = 0.95, $p = 0.037$), and SF12-PCS (adjusted OR = 0.97, $p = 0.021$).

Table 7.7 provides the observed pairwise correlations between variables at different occasions. Though age was a significant predictor of missingness in outcome variables, it showed little correlation with any of the outcome variables. A moderate level of correlation between the outcome variables, except between SF12-PCS and SF12-MCS, was observed on most occasions.

Table 7.7: The observed pairwise correlation between variables

		Age	RMDQ			Pain score			SF12-PCS			SF12-MCS		
			m0	m4	m12	m0	m4	m12	m0	m4	m12	m0	m4	m12
Age		1.00												
RMDQ	**m0**	0.11	1.00											
	m4	0.11	**0.51**	1.00										
	m12	0.20	**0.50**	**0.74**	1.00									
Pain score	**m0**	0.15	**0.62**	**0.39**	**0.40**	1.00								
	m4	0.09	**0.35**	**0.75**	**0.61**	**0.46**	1.00							
	m12	0.15	**0.39**	**0.61**	**0.82**	**0.45**	**0.66**	1.00						
SF12-PCS	**m0**	-0.20	**-0.66**	**-0.45**	**-0.43**	**-0.52**	**-0.33**	**-0.35**	1.00					
	m4	-0.25	**-0.48**	**-0.74**	**-0.67**	**-0.42**	**-0.67**	**-0.58**	**0.58**	1.00				
	m12	-0.28	**-0.47**	**-0.66**	**-0.77**	**-0.43**	**-0.59**	**-0.70**	**0.57**	**0.76**	1.00			
SF12-MCS	**m0**	0.09	**-0.44**	**-0.31**	-0.29	**-0.32**	-0.23	-0.25	0.10	0.22	0.18	1.00		
	m4	0.04	-0.27	**-0.40**	-0.29	-0.20	**-0.31**	-0.24	0.10	0.15	0.16	**0.56**	1.00	
	m12	0.06	-0.29	**-0.37**	**-0.40**	-0.21	**-0.31**	**-0.36**	0.09	0.24	0.18	**0.55**	**0.65**	1.00

m0 - baseline, m4 - month 4 and m12 - month12; Bold values represent the absolute values of the correlations higher than 0.30.

7.6.2 Analysis of STarT Back trial data – estimation of the treatment effect at month 12

7.6.2.1 Results from the actual dataset

Table 7.8 provides the estimates of treatment effect, SE and standardized effect size for the primary and secondary outcomes at the final visit based on a standard ANCOVA (i.e. CCA) and LOCF ANCOVA models. Due to the MDC strategy, the number of participants with complete data on the primary outcome (RMDQ) was substantially higher compared to the secondary outcome measures. Further, a few subjects were excluded from both models for secondary outcomes due to missing baseline data.

Table 7.8: STarT Back - ANCOVA results before and after LOCF imputation of missing values

Outcome variables[1]	Standard ANCOVA[2]				LOCF ANCOVA[3]			
	Estimate	SE	p-value	Standardized effect size[4]	Estimate	SE	p-value	Standardized effect size[4]
RMDQ	-0.964	0.421	0.022	-0.170	-1.010	0.373	0.007	-0.178
Pain intensity	-0.067	0.216	0.757	-0.031	-0.200	0.167	0.231	-0.092
SF12-PCS	1.811	0.835	0.030	0.173	1.963	0.634	0.002	0.188
SF12-MCS	0.983	0.791	0.215	0.083	0.787	0.601	0.191	0.067

Estimate – estimate of treatment effect at month 12 adjusted for age, sex, baseline RMDQ and corresponding baseline of the outcome; SE – standard error of the estimate; [1]RMDQ measured on a 0–24 scale, pain intensity on a 0–10 scale and other variables on a 0–100 scale. [2]Number of subjects included in the analysis was 649 for RMDQ, 541 for secondary outcomes. [3]Number of subjects included in the analysis was 851 for RMDQ, 842 for PRTEE, and 847 for SF12-PCS and MCS. [4]Treatment effect relative to the pooled SD of baseline scores.

Unlike in the TATE trial, the difference in standardized effect size between LOCF ANCOVA and CCA was not substantial in most cases – the largest difference of 0.061 between the two methods was observed for pain intensity. The SE of the estimate of treatment effect was noticeably lower with LOCF ANCOVA compared to CCA in all outcome measures – the largest difference in SEs between the two methods was observed for pain intensity: 0.049 at month 12. However, these deviations in the estimates

and its SE did not affect the overall statistical conclusions for any of the outcome variables.

Table 7.9 presents the treatment effect in the primary and the secondary outcomes at the final visit based on MMRM models. As found in the TATE trial, the MMRM estimates of treatment effect (and hence the standardized effect size) for RMDQ were the same irrespective of baseline scores as a covariate or an outcome. However, a difference was observed for the other variables – pain intensity, SF12-PCS and SF12-MCS – due to missing baseline scores for some participants who responded at later visits. A model with baseline-as-covariate led to the exclusion of the subjects with missing baseline score. The largest difference in the effect size (0.028) was observed between the two baseline handling strategies for pain intensity. Further, the SE of the estimate of treatment effect was slightly lower for the model with baseline-as-outcome for those variables. Therefore, the model with baseline-as-outcome was considered for the remaining analysis in following sections.

Table 7.9: STarT Back - MMRM results

Outcome variables[1]	MMRM (baseline-as-covariate)[2]				MMRM (baseline-as-outcome)[3]			
	Estimate	SE	p-value	Standardized effect size[4]	Estimate	SE	p-value	Standardized effect size[4]
RMDQ	-0.855	0.414	0.039	-0.151	-0.855	0.413	0.039	-0.151
Pain intensity	-0.091	0.209	0.663	-0.042	-0.151	0.208	0.468	-0.070
SF12-PCS	1.873	0.804	0.020	0.179	1.881	0.802	0.019	0.180
SF12-MCS	0.920	0.777	0.236	0.078	1.042	0.775	0.179	0.088

Estimate – estimate of treatment effect at month 12 adjusted for fixed covariates (age, sex, baseline RMDQ for secondary outcomes) and their interaction with time; SE – standard error of the estimate; [1]RMDQ measured on a 0–24 scale, pain intensity on a 0–10 scale and other variables on a 0–100 scale. [2]Number of subjects included in the analysis was 749 for RMDQ, 665 for PRTEE, and 666 for SF12-PCS and MCS outcomes. [3]Number of subjects included in the analysis was 851 for RMDQ, 850 for PRTEE, and 851 for SF12-PCS and MCS. [4]Treatment effect relative to the pooled SD of baseline scores.

For the primary outcome measure (RMDQ), a difference of 0.019 in the standardized effect size between MMRM model and CCA was observed. For pain intensity, it was 0.039

at the endpoint. Whereas for the other remaining secondary outcomes (SF12-PCS and SF12-MCS), the difference in effect size between the two methods was negligible. A marginal reduction in SE with MMRM compared to CCA was found for all the outcome measures.

Table 7.10 presents MI ANCOVA results from the actual dataset. The standardized effect size was generally similar between restrictive and inclusive imputation modelling strategies for all of the outcome measures – the largest difference (0.012) was observed for RMDQ. The SE was also comparable between the two strategies. In addition, the effect size and SE of the estimate of treatment effect were comparable between MMRM with baseline-as-outcome and MI with restrictive imputation modelling.

Table 7.10: STarT Back - ANCOVA results after MI imputation of missing values

Outcome variables[1]	MI (restrictive modelling)[2]				MI (inclusive modelling)[2]			
	Estimate	SE	p-value	Standardized effect size[3]	Estimate	SE	p-value	Standardized effect size[3]
RMDQ	-0.855	0.414	0.040	-0.151	-0.926	0.417	0.027	-0.163
Pain intensity	-0.158	0.211	0.456	-0.073	-0.164	0.204	0.421	-0.076
SF12-PCS	1.874	0.791	0.018	0.179	1.807	0.795	0.023	0.173
SF12-MCS	1.037	0.764	0.175	0.088	1.073	0.792	0.176	0.091

estimate – estimate of treatment effect at month 12 adjusted for age, sex, baseline RMDQ, and corresponding baseline; SE – standard error of the difference; Restrictive modelling – imputation models included only the variables considered for the MMRM analysis models; Inclusive modelling – imputation model for an outcome included, in addition to the variables considered for the MMRM analysis model, other outcome variables as auxiliary variables in order to improve the performance of the imputation procedure; [1]RMDQ measured on a 0–24 scale, pain intensity on a 0–10 scale and other variables on a 0–100 scale. [2]Number of subjects included in the analysis was 851 for all outcome variables. [3]Treatment effect relative to the pooled SD of baseline scores.

Similar to the MMRM results, only small differences in the effect size between MI-restrictive imputation modelling and CCA were found for the outcome variables RMDQ, SF12-PCS and SF12-MCS. For pain intensity, the estimate of treatment effect at month 12 was slightly higher under MI-restrictive imputation modelling, with a difference in effect

size of 0.042. A small reduction in SE with MI compared to CCA was also found for all the outcome measures.

7.6.2.2 Results from the modified dataset and the impact of responses through MDC

Table 7.11 presents what would have been the estimates for the primary outcome (RMDQ) if the MDC strategy had not been employed. This table provides the results from the modified dataset, and shows the standardized effect size and deviation in SE from that observed with the actual dataset. It was found that the standardized effect size obtained from CCA, MMRM and MI with restrictive imputation modelling did not differ in the modified dataset, but the SE of the estimate of treatment effect was slightly lower for MMRM compared to MI and CCA.

Table 7.11: Results from the modified dataset – RMDQ[1]

Results based on	Modified dataset					Deviation in results from actual dataset	
	n	Estimate	SE	p-value	Standardized effect size[4]	SE	Standardized effect size
MMRM[2]	851	-0.812	0.433	0.061	-0.143	0.020	0.008
MI[3]	851	-0.807	0.445	0.070	-0.142	0.030	0.008
ANCOVA	567	-0.818	0.449	0.069	-0.144	0.028	0.026

n – number of subjects included in the analysis; Estimate – estimate of treatment effect at month 12; SE – standard error; [1]RMDQ measured on a 0–24 scale; [2]model with baseline-as-outcome; [3]restrictive imputation modelling. [4]Treatment effect relative to the pooled SD of baseline scores.

The difference in standardized effect size between the actual and modified datasets was trivial for RMDQ score with either the MMRM or MI; however, a larger difference of 0.026 was observed with CCA. The SE of the estimate from all methods increased due to a higher number of missing responses with the modified dataset. Though the estimates showed little difference, the p-values were non-significant for the modified dataset. So the

higher SE values would impact on acceptance/rejection of the null hypothesis, and hence the interpretation of findings.

7.6.3 Summary and interpretation of findings

Unlike in the TATE trial, single imputation of missing responses using LOCF did not make substantial difference to the estimates from CCA of the STarT Back trial data; however, LOCF underestimated SEs compared to the other approaches.

In the actual dataset, both the estimate of treatment effect and SE for the RMDQ were somewhat similar across MCAR-based and MAR-based analyses. Further, the initial exploration of the dropout pattern and predictors of missingness provided evidence against the MCAR scenario in the dataset. However, these findings alone were not sufficient to reject or accept the possibility of MNAR mechanism. The comparison of estimates from the actual dataset (dropout rate at month 12 was 24%) and the modified dataset (dropout rate at month 12 was 34%) found that the RMDQ responses (from 10% of randomized participants) that had been obtained through MDC did not affect the estimate of treatment effect under an MAR-based analysis. That is, the minimally collected responses were potentially ignorable under the MAR-based analysis. Thus, the finding favours a potential ignorable mechanism associated with missing RMDQ responses if the responses that had been retrieved through MDC represent the actual missing data. Therefore, it might be appropriate to conclude that the estimate of treatment effect (and effect size) for the primary outcome under a MAR-based analysis was potentially unbiased.

In the analysis of the STarT Back trial dataset, as observed with TATE trial, it was found that subjects with only baseline data did not make a difference to the estimate of treatment effect at follow-up and there was little reduction in SE by considering the baseline as an outcome with an MMRM analysis. However, implementation of a baseline-

as-covariate model excludes subjects with missing baseline data even though the subjects responded at later visits. This MMRM model yielded a lower estimate of treatment effect and slightly higher SE for the secondary outcomes compared to the model with baseline-as outcome. It was further found that MI ANCOVA by restrictive imputation modelling showed similar results to the MMRM analysis with baseline-as-outcome. In particular, the estimate of treatment effect from the MI ANCOVA was close to that from the MMRM analysis when the number of imputations was increased substantially. In these data, I again used a large number of imputations, which was more than twenty-fold of the current recommendation (White et al., 2011b), and a large number of iterations in the burn-in period, which was fivefold the default value in Stata (StataCorp, 2013), to obtain more accurate estimates of mean difference and SE. MI ANCOVA displayed a slightly higher SE than that from MMRM in the modified dataset where the amount of missing data increased to 33% from 24% in the actual dataset. Further, it was found that the inclusive imputation modelling with MI did not make any noticeable difference to the estimates from restrictive imputation modelling with MI or MMRM analysis even though auxiliary variables showed a moderate level of correlation with outcome variables and association with missingness in these variables.

7.7 Discussion

As noted in chapter 2, any analysis of incomplete data requires unverifiable assumptions about the nature of the missing data, and the validity of inferences from these analyses depends on the correctness of these assumptions. Since it is not possible to verify the correctness of the missing data assumption with certainty based on observed data alone, it has been recommended that one performs sensitivity analyses to assess the robustness of inferences from the primary analysis to a range of alternative plausible missing data assumptions (Food and Drug Administration, 2008; European Medicines Agency, 2010; National Research Council, 2010). Despite this recommendation, sensitivity analyses are

rarely used or reported in practice. The systematic review in chapter 3 found that sensitivity analyses are infrequently and inappropriately used, and insufficiently reported. Even though a fifth (18/86) of trials with missing outcome values at the primary endpoint report having carried out a sensitivity analysis, very few trials (6/18) presented the results of their sensitivity analysis. Either exclusion of subjects with missing data or a single imputation method was the designated sensitivity analysis. Importantly, none of them considered the violation of the MAR assumption. This might be due to the lack of guidance in the literature on how to actually do one. The NRC report on the prevention and treatment of missing data in clinical trial (National Research Council, 2010) admits that there is no established guideline or method in this matter as this is an active area of research. This chapter proposed an approach that makes use of the reminder data (i.e. data that is recovered after number of failed attempts) to assess the influence of missing data on the estimation of treatment effect, and set out to demonstrate this assessment procedure through the empirical evaluation of two incomplete pragmatic trial datasets. This analysis takes a position in between the primary analysis and a detailed sensitivity analysis, and allows a decision to be taken on whether to move ahead with the detailed sensitivity analysis that makes stringent assumptions about missing data.

As pointed out in the background chapter, although some methods (Little, 1988; Diggle, 1989; Ridout, 1991; Fairclough, 2002) have been proposed for the identification of the missing data mechanism, their purpose is generally to detect violations of MCAR assumption by identifying dependence on observed data. Fielding et al. (2009) performed an investigation of missing data mechanisms in a few empirical trial datasets using these previous methods. Similar to the observations from the initial exploration of missing data mechanism in the TATE and STarT Back trials, Fielding et al.'s (2009) investigation found only a distorted view of the mechanism among their datasets. That is, the conclusion about the missing data mechanism was not consistent across the different methods of assessment in their study. The proposed approach in this thesis does not rely heavily on

this kind of initial exploration to investigate the missing data mechanism and to identify the appropriateness of an MAR-based analysis to deal with the missing data.

In accordance with good clinical practice (International Conference on Harmonisation, 1998) and to avoid outcome reporting bias (Dwan et al., 2008), it has been recommended in clinical trials to specify the statistical analysis plan in advance, and it is not advisable to undertake any kind of sensitivity analyses post-hoc. Researchers should pre-specify their plan for handling any potential missing data, to avoid performing several approaches and reporting favourable results. However, it is very challenging to comment on a plausible missing data mechanism and specify a sensitivity analysis to missingess assumptions in advance, with little relevant information in hand. Therefore, it is advisable to consider an MAR-based analysis as the primary analysis method. In this instance, the proposed approach can be used to assess whether a sensitivity analysis to an MNAR assumption should be carried out.

Fielding et al. (2010) introduced an alternative approach to identify which method would be most suitable to deal with missing outcome data in an RCT by utilizing responses later recovered by reminders. A selection of missing data methods were applied to a subset of the actual dataset where individuals with actual missing response at the primary endpoint were excluded. The analyses were repeated for the new reduced dataset but the responses obtained through reminders were regarded as missing. The ‘best’ method that introduced least bias (i.e. difference in estimates between the two new datasets) to the estimation of treatment effect was then suggested as the analysis method most appropriate for the actual dataset. Unless the amount of missing data is minimal, the best method in the new reduced dataset may not be best in the actual dataset because of the exclusion of individuals with actual missing data at the primary endpoint, irrespective of availability of data at an earlier time-point.

Fielding et al. (2012) presented a similar comparison of estimates of treatment effect between the actual data and modified data (reminder responses were regarded as missing) to that provided in the present study. However, the purpose of these comparisons was quite different. In their evaluation, the authors assumed an MAR missing data mechanism that had been assessed in Fielding et al. (2009) and reviewed a number of possible analysis methods (Fielding et al., 2010) using the actual data, and also with the modified data for comparison. The authors did not make use of the findings to validate or invalidate the findings from the actual data; instead, they focused on the difference in the estimate if the reminder strategy had not been implemented. Although missing data in the actual and modified data were 'identified' as MAR, MAR-based analyses – linear mixed-effects model and predictive mean match MI model – resulted in quite different estimates between the actual and modified datasets. They concluded that "*this suggests that ignoring the reminder responses, under-estimates the treatment difference and introduces a bias to the results*". This conclusion has ultimately brought into question their approaches to investigating the missing data mechanism. Under the newly proposed approach, the difference in estimates from the actual and modified data indicates non-ignorability of non-responses in the actual dataset if the reminder responses were representative of non-responses in the actual data.

The newly proposed approach mainly relies on three assumptions: (i) non-responders to the first mail-outs are reminded a number of times before employing an MDC strategy, (ii) data collected via reminder or MDC strategy is treated on an equal footing to that obtained via first mail-out, and (iii) responses that have been retrieved after a number of failed attempts are likely to represent the actual missing responses when the number of failed attempts increases. In pragmatic trials, trialists often follow a strategy of sending a number of reminders to initial non-responders and re-approaching them for minimum data collection, where data collection is usually limited to key outcome measures, if they still have failed to respond. The third assumption becomes more plausible with the minimally

collected responses on an outcome compared to the earlier reminder responses. For example, a subgroup with poor health or disability – common characteristics of participants in a musculoskeletal trial – is highly unlikely to respond to usual mail-outs. The size of a questionnaire and number of repetition (i.e. number of follow-ups) – which usually requires extra effort to complete and return – can de-motivate such participants from responding to the mail-outs; however, these non-responders are possible more likely to respond to a minimum data collection request because they are usually contacted by telephone and it demands only minimal effort on their part. In this example, it is very unlikely to obtain similar estimates from the actual and modified datasets if the third assumption does not hold because a similarity in estimates indicates ignorable MDC responses and MDC responders are unlikely to be better (in terms of improvement in outcome) than non-responders.

The down side of the present approach is that either the reminder or actual missing responses could be non-ignorable when the estimates of treatment effects from the actual and modified datasets are dissimilar and the third assumption does not hold. That is, the dissimilarity of estimates does not confirm the non-ignorability of the actual missing data if the assumption does not hold. Therefore, further sensitivity analysis may be required to assess the impact of departures from an MAR assumption specified with the primary analysis when dissimilarity in estimate of treatment effect between the actual and modified dataset is observed.

Reporting bias – for example, reporting of outcome status at 15 months rather than at the endpoint of 12 months – due to longer follow-up of late responders is a potential limitation of the implementation of reminders and/or minimal data collection approach for reducing the amount of missing data. In addition, the key assumption regarding the representativeness of reminder or MDC responses to the actual missing responses may not be valid if the “true” treatment effect is different at the time of the reminder / MDC

response than at first mail out (due to the reporting bias among late responders). Hence it is important that there is limited influence of time/delay on marginal treatment effect differences in consideration of the proposed approach as a suitable method of evaluating the likely missingness mechanism. In the case of TATE trial, those in the MDC responded a median of 3 months after those who responded to the first mail out at the 12 month follow-up. The observed difference in estimates of treatment effect between the actual and modified datasets indicated the possibility of non-ignorability of the MDC responses, and therefore implied non-ignorability in respect of the missing data. However, the noticeably delayed MDC responses may threaten the validity of the key assumption and the conclusion of non-ignorability of the actual missing responses at the time point of relevance based on the proposed approach unless trialists can ensure lack of reporting bias due to delayed MDC responses.

7.8 Conclusion

In this chapter, I have presented a simple technique to perform a sensitivity analysis to assess the impact of missing data on the primary conclusion using responses subsequently recovered via reminder or MDC. This chapter used two recent pragmatic RCTs to demonstrate this technique. In the TATE trial, this approach showed that an MAR-based analysis is unlikely to yield unbiased estimates of treatment effect for the primary and secondary outcomes (i.e. non-responses were non-ignorable under the MAR-based analysis). Therefore, further sensitivity analyses were required under a range of plausible MNAR assumptions. However, in the STarT Back trial, this approach showed that an MAR-based analysis is unlikely to yield a biased estimate of treatment effect for the primary outcome. In addition, it was found that estimates from MMRM and MI with restrictive imputation modelling were quite similar; however, MI with inclusive imputation modelling resulted in slightly different estimates for some outcome variables and the

estimates further varied by inclusion of additional auxiliary variables within the inclusive imputation model.

Chapter 8: Summary, discussion and conclusions

8.1 Introduction

Randomized clinical trials play a vital role in assessing the efficacy and effectiveness of new interventions compared to a standard or control intervention. Randomization in a clinical trial is intended to generate comparable groups of patients in terms of known and, more importantly, unknown factors that could be associated with the outcome of interest at the onset of the trial. When some outcome measurements are missing, the principal advantage of randomization is threatened, treatment comparisons are potentially biased, and the trial becomes inefficient to detect the treatment effect. Hence, proper treatment of missing data is necessary for a valid analysis; however, it is important to note that none can 'cure' the problem of missing data.

An ITT analysis works well to preserve the benefits of randomization, which is intended to ensure that differences in outcome observed between treatment groups are solely the result of the treatments (Montori & Guyatt, 2001; Heritier et al., 2003), and to reduce the risk of selection bias (Altman, 2009; Fleming, 2011). The ITT principle states that an analysis should be performed by including all study participants in the groups to which they were randomized, regardless of any departures from the original assigned group (Chan et al., 2013). However, the presence of missing data in a trial creates many challenges to implementation of an ITT analysis strategy. White et al. (2012) point out that, for the estimation of treatment effect in clinical trials with missing data, statistical methods that include all randomized participants may sometimes be less valid than methods that do not include all randomized participants, depending on the assumptions made about the missing data. Hence, the researchers state that including all randomized individuals in an analysis of an outcome with missing data is not enough; one should consider an appropriate method to handle the missing data. Further, White et al. (2012)

suggested a framework for an ITT analysis strategy, and this strategy should include a design that is intended to minimize missing data by following up all randomized individuals, an analysis based on a plausible assumption about the missing data, and sensitivity analyses that aim to explore the robustness of the results to a range of alternative plausible assumptions regarding missingness.

8.2 Summary of findings

This thesis consists of three main parts. First, a systematic review has been performed to examine practices relating to methods to handle missing data in published trials in musculoskeletal conditions. Second, a simulation study has been performed to compare the performance of various methods for handling missing data in a longitudinal clinical trial with missing continuous outcome data in a number of scenarios. Finally, an approach has been proposed to investigate the ignorability of the missing data mechanism, and thus to verify the unbiasedness of the estimate of treatment effect from an MAR-based analysis.

8.2.1 Summary of systematic review

The systematic review of RCTs published during January 2010 to December 2011 in five major musculoskeletal journals, which was detailed in chapter 3, identified deficiencies in current practices in relation to dealing with missing outcome data. Many of the reviewed trials failed to obtain outcome data on all randomized participants and/or include all participants in the primary analysis. Almost 95% (86/91) of trials in this review reported dropouts – a high proportion (60%) of trials had more than 10% dropouts and 31% had more than 20%. The high proportion of trials with a significant number of dropouts is of concern.

It was not clear from the reviewed publications whether trialists took careful steps to improve response rate in the trial design and data collection stages. Given that no method can cure the problem of missing data, one should consider ways to try to prevent missing

data during the design and conduct of RCTs. For example, considering sending reminders to encourage participants to respond and limiting data collection to essential variables of interest. Limiting missing data helps to reduce the need to make unverifiable assumptions about the missing data and thus minimizes problems in inferential analyses, especially those problems that flow from misspecification of missing data assumptions in the analyses. The NRC report (National Research Council, 2010) devoted a section of the design aspects of a trial to prevent missing data, and added three recommendations regarding this point (detailed in chapter 3, section 3.6.3).

The present review suggested that published trials continue to use either deletion of cases with missing data or single imputation as the primary approach to dealing with missing data. Nearly 60% (44/75) of reviewed longitudinal trials with missing outcome data used a kind of single imputation to replace missing values – last observation carried forward (LOCF) was the most frequently used single imputation method. Nearly a quarter (18/75) of reviewed trials excluded dropouts who had completed at least one follow-up assessment from the primary analysis. Analyses based on either multiple imputation (MI) or full-information maximum likelihood (FIML) were limited to a small proportion (8/75) of longitudinal trials with missing data.

The review also found that sensitivity analyses are infrequently and inappropriately used, and insufficiently reported. The sensitivity analyses were performed in trials with relatively high proportions of missing data (median 24%; IQR 17%, 33%) but were limited to a low proportion (18/86) of trials with missing data. Furthermore, either exclusion of subjects with missing data (i.e. listwise deletion) or a single imputation method was the designated sensitivity analysis. Very few trials (6/18; 33%) presented the results of their sensitivity analysis, while the others just reported that a sensitivity analysis had been performed and indicated that the findings from the primary analysis were supported by those of the sensitivity analysis. In the present era of the internet, authors have opportunities to publish

sufficient details through online supplements if there is space restriction in the main body of reports.

8.2.2 Summary of simulation study

To address the limitations with the previous simulation studies, which were detailed in chapter 2, and in order to provide a broader and practically more accessible picture of the impact of missing outcome data on estimation of treatment effect in an RCT, a comprehensive simulation study has been used to examine the relative performance of four statistical analysis approaches – CCA, LOCF ANCOVA, MMRM and MI ANCOVA – on a number of possible and credible clinical trial scenarios. The scenarios included various levels of missingness properties (i.e. overall dropout rate, dropout rate between groups, dropout mechanism and the direction of dropouts) along with different levels of data characteristics (i.e. correlation between repeated measurements and data variability, the size of treatment effect, mean trajectory and sample size). The study methodology was specified in detail in chapter 4: incomplete data were generated with pre-specified dropout rates (equal and differential dropout rates between groups) under different missing data mechanisms: MCAR, MAR-B, MAR-L and MNAR. The MAR and MNAR dropout mechanisms were implemented under two contrasting scenarios. In the first scenario, dropouts were in the same direction in both study groups (MAR-B1, MAR-L1 and MNAR-1) – dropouts were a random sub-sample of subjects who did poorly in both study groups. Results from an opposite scenario, wherein dropouts were a random sub-sample of subjects who did well in both study groups, are provided in appendix 4; as expected results were the same as in the first scenario but differed in respect of the direction of bias. In the second scenario, dropouts were in opposite directions between study groups (MAR-B2, MAR-L2 and MNAR-2) – dropouts were a random sub-sample of subjects who did poorly in the control group and those who did well in the experimental group. Results from an opposite scenario wherein dropouts were a random sub-sample of

subjects who did well in the control group and those who did poorly in the experimental group are provided in appendix 4; as expected, results were the same as in the second scenario but differed in respect of the direction of bias. For any analysis of incomplete data in practice, it is necessary to make assumptions, which are often unverifiable in the incomplete data, about the missingness. It is crucial, therefore, to assess the performance of methods to deal with missing data in relation to a variety of contrasting scenarios in order to understand the robustness of the results under the different missing data handling approaches to variation and possible extreme situations in the missing data mechanism and criteria for dropout, and to aid interpretation of findings from an incomplete dataset.

Table 8.1 presents a summary of key simulation results that were reported in chapter 5 in which the study assessed the relative performance of the missing data methods with respect to bias, CI coverage and statistical power in relation to the estimation of treatment effect.

Table 8.1: Summary of simulation results*

Method	Acceptability[1]	Missing data mechanism	Absolute bias in standardized effect size†‡		Coverage (%) of 95% CI		Statistical power (%)	
			10% dropout rate	30% dropout rate	10% dropout rate	30% dropout rate	10% dropout rate	30% dropout rate
CCA	Acceptable	MCAR; MAR-B1; MAR-B2; MAR-L1 (with equal dropout rate between arms); MNAR-1 (with equal dropout rate between arms)	0.000–0.021	0.000–0.018	94.0–96.9	93.3–96.6	82.5–88.6	64.9–80.6
	Unacceptable	MAR-L1 (with differential dropout rate between arms); MAR-L2; MNAR-1 (with differential dropout rate between arms); MNAR-2	0.022–0.267	0.078–0.529	68.0–95.8	5.7–94.5	38.3–97.0	4.9–99.7
LOCF	Unacceptable	All	0.000–0.306	0.015–0.634	67.2–98.0	6.9–99.1	47.6–99.9	4.5–99.9
MMRM	Acceptable	MCAR; MAR-B1; MAR-B2; MAR-L1; MAR-L2; MNAR-1 (with equal dropout rate between arms)	0.000–0.020	0.000–0.022	93.6–96.4	93.2–96.3	84.0–89.2	68.3–81.3
	Unacceptable	MNAR-1 (with differential dropout rate between arms); MNAR-2	0.040–0.205	0.138–0.462	73.1–94.2	11.4–91.3	45.5–96.6	4.6–99.5
MI	Acceptable	MCAR; MAR-B1; MAR-B2; MAR-L1; MAR-L2; MNAR-1 (with equal dropout rate between arms)	0.000–0.020	0.000–0.022	93.6–96.4	93.8–97.3	82.2–87.5	64.3–79.8
	Unacceptable	MNAR-1 (with differential dropout rate between arms); MNAR-2	0.041–0.205	0.138–0.462	73.8–94.6	12.3–92.8	44.0–96.5	4.4–99.5

*126 scenarios with 10% and 30% dropout rate; †effect size standardized for baseline data variability; ‡the bias in standardized effect size ranged 0.001–0.011 with no missing data; [1]a method is considered as acceptable if the standardized effect size did not vary by dropout rate.

In the simulation study, LOCF yielded biased estimates of treatment effect in most scenarios irrespective of missing data mechanisms, and there was no consistency in the magnitude of bias across the considered scenarios. For example, LOCF overestimated the treatment effect for scenarios of higher dropout rate in the control group under MCAR and MAR mechanisms, but underestimated the treatment effect in the same scenario under an MNAR mechanism. Furthermore, the bias varied by dropout rates, the direction of dropout, the timing of dropout, data variability, and the level of correlation between repeated measurements. Even with 10% dropout rate and an MCAR mechanism, the magnitude of the bias exceeded 23% of baseline SD for a scenario in which there was higher dropout rate in the experimental group. With LOCF, the statistical power in some scenarios was close to 100% due to overestimation of treatment effect and underestimation of SE.

In the present simulation study, CCA yielded unbiased estimates of treatment effect for scenarios of equal dropout rate and the same direction of dropouts in both treatment groups, irrespective of missing data mechanism. Furthermore, the analysis also produced unbiased estimates of treatment effect under MCAR and MAR-B (MAR dependent on the baseline), irrespective of other scenarios. In all these scenarios, CCA maintained the CI coverage at an acceptable level, but not the statistical power. The loss of statistical power with 10% dropout rate was 2%–7% for MCAR, 2%–6% for MAR-B1 (same direction of dropout), and 3%–8% for MAR-B2 (opposite direction of dropout) depending on various levels of data variability and correlation between the baseline and the final endpoint. The effect of differential dropout rate and differential direction of dropouts on statistical power were not noticeable at 10% dropout level but were so somewhat at 30% dropout level. A CCA estimate of treatment effect was biased under MAR-L and MNAR mechanisms, except in the scenario of equal dropout rate and same direction of dropouts in both groups. With 10% dropout rate, the bias was up to 13% of baseline SD under MAR-L and

27% of baseline SD under MNAR mechanism depending on other considered factors. Correspondingly, the CI coverage and power were also affected under the MAR-L and MNAR mechanisms.

MMRM and MI-based analyses produced unbiased estimates of treatment effect in all scenarios in which CCA yielded unbiased estimates. These analyses also produced unbiased estimates under MAR-L. In all these scenarios, CI coverage was maintained but statistical power was not. Both methods yielded similar estimates of treatment effect in all scenarios. However, a slightly lower coverage (maximum difference of 1% for 10% dropout rate) and higher statistical power (maximum difference of 3% for 10% dropout rate) was observed with MMRM compared to MI, especially under scenarios of differential dropout rate and opposite direction of dropouts. Furthermore, the difference slightly increased when the dropout rate was increased from 10% to 30%. Importantly, the loss of power in these methods was not substantially different from CCA in scenarios in which the estimate of treatment effect was unbiased. Estimates of treatment effect from MMRM and MI-based analyses were biased under the MNAR mechanism, except in the scenario of equal dropout rate and same direction of dropouts in both groups. With 10% dropout rate, the bias was up to 12% of baseline SD under MNAR-1 and 20% of baseline SD under MNAR-2, depending on other considered factors. Correspondingly, the CI coverage and power were also affected under the MNAR mechanisms.

8.2.3 Summary of empirical evaluation

Chapter 7 presented re-analyses of two pragmatic RCTs – TATE and STarT Back trials – that included a reminder process for non-responders. The dropout rate in respect of the primary outcome at the final visit was 26% for the TATE and 24% for the STarT Back. Both trial datasets had been previously analysed using MI with an inclusive imputation

modelling strategy (i.e. a strategy that included all variables[17] in a dataset in the imputation model). The present study re-analysed these datasets using standard ANCOVA, LOCF ANCOVA, MMRM and MI ANCOVA (with restrictive imputation modelling strategy). In the TATE trial, it was found that all but LOCF-based analyses of the primary outcome yielded a similar standardized effect size, which was equivalent to the estimate from the original analysis. The LOCF-based analysis led to double the estimate. In the STarT Back trial, the estimates of treatment effect were not consistent across the analysis methods – CCA and LOCF slightly overestimated the treatment effect compared to the estimates from MMRM and MI. Additionally, the estimates from all these methods were slightly lower than the estimate reported from the original analysis but the p-values remained less than 0.05 in all the analyses. It was observed that the estimates of treatment effect from MI varied by use of inclusive imputation modelling strategies with different number and types of auxiliary variables.

A comparison of MMRM estimates of treatment effect from the actual and modified (responses from minimal data collection (MDC) were regarded as missing) datasets found a difference between these estimates in the TATE trial but not in the STarT Back trial. This finding implies that the MDC responses had an impact on the estimate of treatment effect in the TATE trial, and these additional responses should not be ignored from the estimation. If it can be assumed that the MDC responses were representative of the non-responses, the MMRM estimate of treatment effect from the TATE trial was more likely to be biased. Therefore, further sensitivity analyses to assess the robustness of the estimate of treatment effect were required in the TATE trial under a range of plausible MNAR assumptions. Judging from the change in direction of effect, an analysis addressing

[17] Although the original dataset involved a substantial number of variables, the re-analysis used only six variables (a primary outcome variable, three secondary outcome variables, and two additional baseline variables (age and sex)).

MNAR with missing data mimicking MDC responses would result in a narrowing in the effect estimate between treatment arms (the same conclusion would likely hold). By contrast, the MMRM estimate of treatment effect from the STarT Back trial was fairly consistent between actual and restricted datasets – so the original MAR evaluation is unlikely to be biased if the assumption holds that late responders typify non-responders.

8.3 Discussion of the findings

8.3.1 The performance of incomplete data analysis methods for the estimation of treatment effect in RCTs

8.3.1.1 Analysis using last observation carried forward

The present simulation study showed that LOCF ANCOVA is very unlikely to give an unbiased estimate of treatment effect. This study further found that LOCF does not yield a better estimate of treatment effect – in terms of bias – compared to CCA, MMRM or MI ANCOVA when the missing data mechanism is MCAR or MAR. Since values imputed with LOCF are treated as 'true' observations and do not add any sort of component of uncertainty in the estimation, this approach generally underestimates the SE of the treatment effect as evident in the simulation study. The biased estimates of treatment effect and SE introduce artificial inflation or deflation of CI coverage and statistical power, depending on the direction and magnitude of the bias. The findings clearly contradict the statement that LOCF-based analysis relies on an MCAR assumption (Mallinckrodt et al., 2001a; Mallinckrodt et al., 2008; Fielding et al., 2010; Bredemeier, 2012) – this assumption is, therefore, unwarranted.

In the simulation under an MNAR mechanism, LOCF was more favourable compared to other missing data approaches in a few circumstances. Lower bias for LOCF was observed when there was higher dropout rate in the control group in the presence of a greater improvement in outcome in the intervention group. Further, the simulation study

(Chapter 6) that assessed the robustness of the findings to variations in trajectory profile found that the performance of LOCF varied considerably across the scenarios, especially when dropout rate was higher in the intervention group with substantial improvement in outcome. It appears that differential improvement in outcome between treatment groups and the timing of dropout substantially affect the estimation of treatment effect using LOCF-based analysis. Hence, an evaluation of trajectory profile by dropout patterns may be helpful in assessing the impact of the LOCF approach.

One argument in favour of LOCF is that an LOCF-based analysis does not bias the treatment effect in favour of a new intervention against a control intervention (Streiner, 2008; Navarro-Sarabia et al., 2011). Results from previous simulation studies by Baron et al. (2008) and Olsen et al. (2012) seem to support this argument. Their work on an MNAR mechanism showed that LOCF led to an underestimation of treatment effect with reduced (and hence, conservative) statistical power for all considered scenarios. Their finding might be true with a trajectory profile that had been used for their simulation studies. However, as found in the present simulation study, this may not always be true, since overestimation of treatment effect is evident when there are more or earlier dropouts in the control group than in the intervention group and the rate of improvement on an outcome in the intervention group is higher than in the control group. In a previous study (Mallinckrodt et al., 2001a), the authors reported that LOCF-based analysis overestimated the true effect in treatment versus ineffective placebo comparisons and underestimated the true treatment effect in treatment versus effective placebo comparisons. The present study identified that the direction of bias with LOCF can also be influenced by many other factors, such as the timing of dropout, direction of dropout, and differential dropout rate between treatment groups. In the TATE trial (Chapter 8, section 7.3), the estimates of treatment effect in the primary (pain intensity) and the secondary (PRTEE) outcomes from LOCF ANCOVA were higher than those of CCA, MMRM and MI ANCOVA. It seems the magnitude of the bias in estimate of treatment effect from an LOCF-based analysis is

quite unpredictable in any scenario. Some researchers may use LOCF to maintain the sample size. This ultimately underestimates the SE of treatment effect in most situations, and therefore ends up with an unacceptable CI coverage.

In summary, the use of LOCF-based analysis as the primary analysis in an RCT should be avoided. A very recent search of PubMed for of "*("LOCF"[All Fields] OR "last observation carried forward"[All Fields]) AND (clinical trial[ptyp] AND ("2014/01/01"[PDAT] : "2014/09/30"[PDAT]))*" resulted in 15 hits. Even though the lower number is promising, it cannot be concluded that LOCF-based analysis has been completely discarded.

8.3.1.2 Analysis of covariance without imputation of missing values (Complete-case analysis)

A standard ANCOVA yields an unbiased estimate of treatment effect under MCAR or MAR dependent on the baseline (MAR-B) – which is sometimes referred to as covariate-dependent MAR – irrespective of dropout rate, direction of dropout, data variability and correlation between repeated assessments. So, a standard ANCOVA analysis ensures an unbiased estimation of treatment effect if the missingness in outcome is truly associated with the covariates in the ANCOVA model. In addition, CCA retains the targeted CI coverage to an acceptable level in all these scenarios. However, the width of CI increases with decreases in the number of complete cases, and the width further varies by the direction of dropout (whether dropouts are in the same direction in both groups or not) – a slightly wider CI is obtained for scenarios whereby dropout is in the opposite direction between treatment groups. The changes in CI width directly translate to loss of statistical power.

Many authors in the systematic review highlighted an equality in dropout rate between treatment groups. The general view appeared to be that equal dropout rate between groups would not lead to a biased estimate of treatment effect. This is true with CCA if the

MCAR or MAR-B assumption is satisfied. In addition, this is also true for MAR dependent on the last observed values (MAR-L) – also referred to as outcome-dependent MAR – and MNAR data if dropouts are in the same direction in both groups. That is, in all situations where dropout rates are equal between the groups and dropouts are in the same direction, CCA produces unbiased estimates of treatment effect with an acceptable CI coverage irrespective of missing data mechanism. Bell et al. (2013) reported a similar conclusion based on a simulation study with a 30% dropout rate. In the present study, 'same direction' means dropouts performed either well in both groups or worse in both groups, and 'opposite direction' means dropouts in one group performed well and dropouts in the other group did worse. Intuitively, the reality is that missingness is unlikely to be completely skewed between treatment groups (i.e. entirely in opposite direction) and is likely to be somewhere in between. Thus, the findings serve to caution against a conclusion of unbiasedness just because of equal dropout rate between treatment groups – unless it can be justified in some way that dropout in the same direction is the most plausible explanation for missingness.

For all MAR-L and MNAR scenarios except the scenario of equal dropout rate with the same direction of dropout, CCA analysis gives biased estimates of treatment effect. In the present simulation study, it was observed that the bias ranged from 2% to 27% of baseline SD (Table 8.1) for scenarios of a 10% dropout rate and 8% to 53% of baseline SD for scenarios of a 30% dropout rate. As expected, bias in estimate of treatment effect from CCA is severe for MNAR scenarios with opposite direction of dropout, and is greater than for MAR-L missingness. In particular, even with 10% dropout rate this analysis method can lead to substantially biased estimates of treatment effect. In the present study, the treatment effect was underestimated in most MAR-L and MNAR scenarios, especially for scenarios with opposite direction of dropout, since individuals with favourable responses in the intervention group and individuals with unfavourable responses in the control group

were regarded as the dropouts for the scenario with the opposite direction of dropout. In general, the bias adversely affects the desired CI coverage and statistical power.

Although a CCA produces less biased estimates of treatment effect in many scenarios compared to a LOCF-based analysis, the CCA does not consider the availability of outcome responses at interim visits. In many situations in clinical trials, these interim measurements have some level of influence on participants' decision to continue a trial (Prakash et al., 2008). By ignoring these interim measurements CCA is omitting available and potentially informative data in respect of the treatment effect and thus leads to biased estimation of treatment effect. Therefore, CCA should be disregarded in favour of a more efficient analysis method that can effectively utilize the additional available data, irrespective of the amount of missing data.

8.3.1.3 Mixed-effects model for repeated measures and MI-based analysis of covariance

The estimate of treatment effect at the final time-point from MMRM is based on the FIML approach and is implicitly adjusted for the outcome observed at the previous time-points and their correlation with the final time-point (Davis, 2014). By contrast, MI requires explicit imputation of missing values and involves multiple random draws from the posterior predictive distribution of the missing data under a posited Bayesian model, instead of a single set of FIML estimates of parameters (Ratitch, 2014).

In the present simulation study, MMRM and MI ANCOVA are found to be more robust to bias from missing data compared to CCA and LOCF ANCOVA. That is, MMRM and MI-based analyses produce valid estimates of treatment effect for MCAR, covariate-dependent MAR and outcome-dependent MAR (MAR-L) scenarios, irrespective of dropout rate and direction of dropout. Previous studies (Lane, 2008; DeSouza et al., 2009; Siddiqui et al., 2009; Siddiqui, 2011; Bell et al., 2013) that evaluated a number of dropout

scenarios under an MAR dropout mechanism also confirmed that MMRM performed well – in terms of controlling bias – compared to LOCF and CCA. However, the present study finding on MI ANCOVA contradicts the results from a previous study by Siddiqui (2011), who reported a biased estimate of treatment effect in relation to a scenario of outcome-dependent MAR. In addition, the present simulation study shows that both methods yield unbiased estimates of treatment effect in the MNAR situation whereby dropout is equal and in the same direction between treatment groups. This result (in the context of MMRM) agrees with previous work by Bell et al. (2013).

Siddiqui (2011) also reported MI-based analysis as being a conservative approach with type 1 error rate of 1% and a high SE (compared to MMRM) in an MAR scenario of differential dropout rate. However, the present study countered his finding by virtue of the fact that both methods (MMRM and MI ANCOVA) attain the accepted level of CI coverage (i.e. 93.6%–96.4%) for all MCAR and MAR data scenarios, irrespective of dropout rate. A slight difference in CI coverage and width between these methods was observed in the present study. This difference was most noticeable for scenarios with opposite direction of dropout – MI retained very slightly higher CI coverage and width than MMRM. The present study also found that data variability had a slightly greater impact on the width in respect of MI-based analysis compared to MMRM. In this simulation, only a limited number of imputations – a number equivalent to percentage of dropouts – was used as recommended by White et al. (2011b). However, it is noted that a substantial number of imputations are required to minimize the variability due to randomness associated with MI (Rubin, 1987; Graham et al., 2007).

In the empirical evaluation of the TATE and STarT Back trials, it was found that a large number of imputations need to be used with MI to minimize the difference in estimate of treatment effect between MI and MMRM analyses. In the TATE dataset, a large number of imputations, which was more than ten-fold of the current recommendation (White et al.,

2011b), was used to obtain a closer estimate from the two analysis methods. Similarly, in the STarT Back data, the number of imputations used in MI was more than twenty-fold of the current recommendation (White et al., 2011b). However, the SE with MI ANCOVA was still slightly higher than that from the MMRM analysis. This confirms the finding from previous studies (Collins et al., 2001; Schafer & Graham, 2002; Barnes et al., 2008) that FIML-based methods, in general, produce slightly smaller SEs than MI-based methods unless the number of imputation is substantially high.

Under all MNAR data scenarios, except the scenario of equal dropout rates with the same direction of dropout, none of the considered approaches performs well in terms of controlling bias in estimation of treatment effect; the bias markedly increases in relation to an increase in overall dropout rate and data variability. Further, both MMRM and MI yield similar estimates of the treatment effect and RMSE in all scenarios. In the present simulation study, it was observed that the bias ranged from 4% to 21% of baseline SD (Table 8.1) for the scenarios with 10% dropout rate and 14% to 46% of baseline SD for scenarios with 30% dropout rate. It was further found in the present study that the bias is severe for scenarios of 'opposite direction' of dropout compared to 'same direction' of dropout when the missing data mechanism is MNAR. Moreover, the bias can be substantial with only a 10% dropout rate. However, estimates are slightly more appropriate than those given by CCA in some instances, especially for scenarios of strong correlation between repeated follow-ups. Further, the bias adversely affects the desired CI coverage and statistical power. Previous studies by Mallinckrodt et al. (2001b; 2001a; 2004), in which data were simulated under an MNAR mechanism and a high overall dropout rate, showed only a negligible bias in MMRM estimates of treatment effect. Another simulation study by Barnes et al. (2008) also showed that MMRM and MI retained an acceptable level of CI coverage under similar scenarios of MNAR mechanisms. However, the present study provides a clear warning against the use of MMRM or MI-

based analyses when an MNAR missing data mechanism is suspected. Lane (2008) also showed some scenarios in which MMRM can be severely flawed.

The present study is partially comparable to the previous simulation study by Olsen et al. (2012), in which they compared CCA, MMRM and MI ANCOVA for the analysis of a continuous outcome variable in MNAR scenarios of an approximately 30% overall dropout rate. For an equal dropout rate scenario, the previous study reported valid (with no/negligible bias) estimates of treatment effect from these three methods. However, the present study identified, in line with Bell et al. (2013), that the implication of validity cannot be generalized to all equal dropout scenarios, and that the only possible scenario for an unbiased estimate under an MNAR mechanism is the scenario of equal dropout rate with 'same direction' of dropout. For a differential dropout rate scenario, Olsen et al. (2012) reported biased estimates from these analysis methods, but lower bias with MMRM compared to CCA and MI ANCOVA. The present study also confirms the possibility of bias from these methods for the same scenario, but does not confirm the superiority of MMRM over MI ANCOVA in terms of bias. The previous study also reported that CCA yielded a high type 1 error rate in a scenario of differential dropout rate and acceptable level of type 1 error rate in a scenario of equal dropout rate. The present comprehensive simulation study (with 546 scenarios in chapter 5 and 6) concluded that all methods retain the acceptable CI coverage (which can be directly translated to acceptable type 1 error rate) when the estimate of treatment effect is unbiased.

8.3.2 Choice between MMRM and MI-based analyses in an RCT

There are a number of practical difficulties associated with MI implementation. One disadvantage with MI is the inconsistent estimates due to random draws; a high number of imputations are required to reduce the variability due to randomness (Graham et al., 2007). Depending on the number of imputations and the model specifications to impute missing values, more computing time might be required. Since only a limited number of

variables were involved in the simulation study (i.e. an outcome variable and treatment indicator), the implementation of MI was as simple as the MMRM model. However, in practice where an analysis model involves a number of covariates and interactions, it is quite difficult to specify an imputation model in MI since the imputation model should be congenial (Molenberghs & Kenward, 2007). That is, the imputation model should at least include all the variables, interactions and transformations (e.g. non-linear terms) that are also intended for use in the analysis model. In the simulation study, a 'restrictive' imputation model (i.e. the model included only those variables that used for the MMRM model) was used. From the simulation and empirical evaluation of MMRM and MI ANCOVA, it is clear that both methods yield similar estimates of treatment effect and SE if the number of imputations is substantially high. That is, as Rubin (1987) showed, an infinite number of imputations are required to make MI ANCOVA as efficient as MMRM. However, with the recommended (White et al., 2011b) number of imputations, it is possible to get sufficient efficiency for MI, and the difference in SE between the two methods is very unlikely to influence statistical decision-making. In summary, there is no advantage of using MI with restrictive imputation as opposed to MMRM unless covariates have missing values.

Further, it is possible to incorporate auxiliary variables that are not part of a subsequent analysis model into the imputation model in order to make MAR more plausible and, therefore, to increase efficiency and reduce bias (Collins et al., 2001; Spratt et al., 2010) – this is referred to as an inclusive modelling strategy (Collins et al., 2001). This strategy is generally implemented in two ways: (i) include all auxiliary variables into the imputation model or (ii) include variables that are selected based on data considerations. Collins et al. (2001) performed a simulation study in which they compared the restrictive and inclusive imputation modelling strategies to assess the influence of auxiliary variables on the estimation of population mean and regression parameters in a linear regression model on an outcome variable with missing values. Referring to Collins et al.'s (2001) work, the

advantage of auxiliary variables is considerable in estimation of treatment effect only if those variables are strongly associated with missingness in the outcome and highly correlated with the observed outcome. Further, the inclusion of too many 'junk' variables (variables that are not associated with either missingness or the observed outcome) may increase bias in estimation of treatment effect and decrease precision (Collins et al., 2001; Hardt et al., 2012). Spratt et al. (2010) conducted an empirical evaluation of inclusive imputation models using a longitudinal observational dataset. In that dataset, the outcome variable (wheeze at age 81 months) and two of three prognostic variables were subject to missing data. They reported that inclusion of only those auxiliary variables that were predictive of variables having missing data increased odds ratio and reduced SE. However, the inclusion of only those variables associated with missingness did not make a difference to the OR and SE. The impact of 'junk' variables was not reported in their study. Enders (2010) suggested selecting auxiliary variables that have correlations greater than ±0.4 with variables with missing data. In line with Collins et al.'s (2001) work, White at al. (2011b) pointed out that one should include variables that are associated with the missing data mechanism and/or correlated with the variables having missing observations. However, Thoemmes and Rose (2014) argue that there are auxiliary variables that can induce bias in estimation using MI irrespective of the association between the auxiliary and outcome variables. In general, there is no consensus on the selection of the auxiliary variables. Importantly, if researchers consider the inclusion of auxiliary variables that are selected based on data considerations as suggested by White at al. (2011b), the analyses of primary and secondary outcome variables in an RCT data may require separate imputation models, costing considerable effort and time. Therefore, further research is necessary to develop guidance on selecting auxiliary variables in MI for the analyses of longitudinal RCTs where there are several baseline, secondary outcome and primary outcome variables, and where both primary and secondary outcomes have missing responses.

As noted, the present simulation study did not address the inclusive MI modelling strategy. However, comparisons of restrictive and inclusive imputation modelling with MI were performed on a range of outcome variables from the TATE and STarT Back trial datasets. In the TATE trial, it was found that the inclusive imputation modelling with MI resulted in slightly different estimates of treatment effect and SE from restrictive imputation modelling with MI or MMRM analysis. The observed difference varied by the number and type of auxiliary variables added into the inclusive imputation model even if the auxiliary variables were not associated with an outcome variable and its missing indicator. For example (appendix 8), the effect size for the secondary outcome PRTEE was −0.215 based on MI with restrictive imputation modelling strategy. The inclusion of pain intensity (for which additional observations were available due to the MDC strategy) as an auxiliary variable in the imputation model led to a reduced effect size of −0.193; whereas the inclusion of SF 12-PCS led to an increased effect size of −0.261. The inclusion of both auxiliary variables in the restrictive imputation model for PRTEE yielded an effect size of −0.212. In the TATE trial dataset, pain intensity showed a strong correlation with the PRTEE but not the SF 12-PCS. In the STarT Back trial, it was found that the inclusive imputation modelling with MI did not make any noticeable difference[18] to the estimates from restrictive imputation modelling with MI or MMRM analysis even though included auxiliary variables showed a moderate level of correlation with outcome variables and moderate association with missingness in the outcome variables. These findings make a strong case for further research on selection of auxiliary variables in missing data problems.

[18] However, inclusion of all variables in the original dataset (where the number of variables is substantially greater than that in the dataset used for the re-analysis) showed a difference in estimate from the original analysis published.

In summary, unless a researcher is more comfortable with MI than with mixed models, it is better to utilize an MMRM model as the primary analysis method. However, MI with an inclusive modelling strategy can be an ideal starting point for a sensitivity analysis.

8.3.3 Strategy for handling baseline values with MMRM analysis

In chapter 6, comparisons of the results from MMRM with baseline-as-covariate to an alternative model with baseline-as-outcome were reported. Since the presence of early dropouts is common in pragmatic trials, all simulated datasets in this thesis involved some participants without any follow-up measurements. Hence, analysis using MMRM with baseline-as-covariate led to the exclusion of those participants from the analysis. In the present simulation study, it was found that the inclusion of participants without any follow-up data (by considering baseline as an outcome) did not make a difference to the estimates of treatment effect. However, with a 30% dropout rate (nearly 10% were early dropouts), the study found a slight difference in the coverage of the 95% CI and the observed power due to slightly reduced average SEs for scenarios of opposite direction of dropout. The model with baseline-as-covariate and Kenward-Roger correction for a finite sample showed the highest coverage but the lowest statistical power, whereas the model with baseline-as-outcome and without the correction showed the lowest coverage but the highest power. Since it was found that the difference is only a concern for some extreme scenarios of dropout, the choice of baseline handling strategy is unlikely to influence the CI coverage or statistical power when baseline is fully observed. Thus, the findings of the present study do not concur with previous work by Lui et al. (2009) in which the authors found that retaining baseline as a covariate could result in greater loss of efficiency compared to the other approach and hence favoured the model with baseline-as-outcome. The analysis of the TATE and STarT Back trial datasets (Chapter 7) also confirmed that the inclusion of participants without any follow-up data in MMRM analysis (by considering baseline as an outcome) does not make a difference to the estimates of treatment effect

at the endpoint, but there was a small reduction in SE. Overall, the present study supports Kenward et al.'s (2010) disagreement with Lui et al.'s (2009) conclusion. Since both the strategies do not take into account early dropouts in the estimation of treatment effect, it is consequently important to ensure that these early dropouts do not have any adverse impact on the benefit achieved through randomization.

In a setting with missing baseline data, the implementation of MMRM with baseline-as-covariate excludes individuals with missing baseline data. In the STarT Back trial it was observed that this MMRM model yielded a lower estimate of treatment effect and slightly higher SE for the secondary outcomes compared to the model with baseline-as-outcome. However, it is also possible to retain the same power using MMRM with baseline-as-covariate without the risk of bias using supplementary methods i.e. mean imputation or missing indicator method (White & Thompson, 2005).

8.3.4 The benefits of sample size inflation to the effect of attrition on statistical power

The effect of missing data on the statistical power to detect the true treatment effect in a clinical trial has been detailed in chapter 5 and 6. The present simulation study has established that trials could be artificially underpowered/overpowered depending on the magnitude and direction of bias in estimates. It was also found that trials are underpowered even with unbiased estimates of treatment effect irrespective of analysis methods or missing data mechanisms. When there is no bias in the estimate of treatment effect, the reduction in power is shown to be associated with the direction of dropout, the amount of missing data, data variability and correlation between repeated measurements. The reduction is relatively larger in scenarios of 'opposite direction' of dropouts between study groups compared to the scenarios of 'same direction' of dropouts in both groups.

Though previous simulation studies (Mallinckrodt et al., 2001a; Baron et al., 2008; Lane, 2008; DeSouza et al., 2009; Siddiqui, 2011; Olsen et al., 2012) have shown the effect on power of the presence of missing data in a few scenarios, these studies were generally not sufficient to extend their inference into real practice. This is mainly because these simulation studies used unrealistic sample sizes that yielded substantially higher or lower power in the absence of missing data than the routinely desired power of 80% or 90% in real clinical trial scenarios. Unlike the previous studies, the present study in chapter 5 used sample sizes that were calculated in order to ensure 90% power for data without missing values and the effects of missing data were explored in contrasting scenarios for the purpose of generalizability.

In the literature, discussion about the loss of efficiency in terms of power is generally limited to the methods that lead to listwise deletion of cases with missing values. Thus, researchers may be less concerned over the possible effect of missing data on the desired power, particularly when the planned analysis is other than CCA. For example, in a recently published statistical analysis plan for a trial (Johnsen et al., 2014), the authors reported that an MI-based approach will be used as the primary analysis, and they claimed that their previous experience supports the MAR assumption. The authors reported the sample size calculation in detail, but they failed to comment on loss in power due to an expected attrition in their calculation. However, they had targeted a 90% statistical power in the sample size calculation. The consideration of a high power might be because of expected dropouts but the report did not detail any such explanation. The systematic review in chapter 3 also found that many of the reviewed trials did not consider the attrition effect on statistical power in their sample size calculation. In the systematic review, it was found that 21% of trials (19/91) failed to report a formal sample size calculation, contrary to CONSORT recommendations (Moher et al., 2010). Furthermore, many of the trials (44/72) that reported a sample size calculation did not address the adjustment for anticipated dropout rate. This indicates a need to promote the importance

of adjustment for attrition in sample size calculations in order to retain sufficient statistical power to detect a true treatment effect. Importantly, the present simulation study shows that none of the considered methods is immune to loss of power due to dropouts.

An extract of simulation results regarding observed statistical power under the WM[19] variance-covariance scenario is provided in table 8.2. The sample size of 75 per group (columns two and four) was calculated to detect the true difference of −9.0 with 90% power in the absence of missing data. These columns provide the range of statistical power for each analysis method among the scenarios of unbiased estimates of treatment effect. For example, the observed power ranged from 84.1% to 87.7% for CCA when the dropout rate was 10% and the estimates were unbiased. It can be seen that all three methods almost equally underestimated the nominal power (even with 30% dropout rate). As noted earlier, the slightly lower power with MI ANCOVA can be improved by considering a substantially larger number of imputations. The lower limits of the results (i.e. larger underestimation of power) in each method were associated with scenarios for opposite direction of dropout with differential rate of dropout, and in contrast the upper limits of the results were associated with scenarios of same direction of dropout with equal dropout rate between groups. As was seen, the amount of missing data and the direction of dropout are the two major factors that contribute to the underestimation of power (due to an increase in average SE).

[19] Weak correlation with moderate SD

Table 8.2: Statistical power for analysis methods (among the scenarios in which the methods yielded unbiased estimates of treatment effect)

Analysis method	10% dropout rate		30% dropout rate	
	n=150[1]	n=168[2]	n=150[1]	n=216[3]
CCA	84.1%–87.7%	88.7%–90.3%	66.0%–78.9%	82.6%–92.0%
MMRM	84.6%–87.6%	88.8%–90.0%	68.3%–79.8%	83.2%–92.1%
MI ANCOVA	83.1%–87.5%	87.7%–90.0%	65.8%–77.5%	81.4%–91.6%

Percentages are range of power among the scenarios in which the methods yield unbiased estimate of treatment effect; [1]sample size calculated to detect the true difference of −9.0 with 90% power in the absence of missing data; [2]adjusted for 10% dropout rate; [3]adjusted for 30% dropout rate.

In chapter 6, an additional simulation study was performed to evaluate the common practice of inflating sample size – by the inverse of one minus the anticipated dropout rate – for the purpose of retaining the desired power at a nominal level of 80% or 90% in the presence of dropouts. Overall, the inflation in sample size was helpful in protecting against the loss of power due to attrition. For example, columns three and five in table 8.2 provide the range of observed statistical power (desired power was 90%) for each analysis method among the scenarios of unbiased estimates of treatment effect with increased sample size for dropout rate of 10% and 30%, respectively. In the case of 10% dropout rate, the methods – CCA, MMRM and MI – could retain the observed power very close to the desired level with the increased sample size in dropout scenarios of unbiased estimate of treatment effect. In the case of 30% dropout rate, these methods could also retain the observed power to the desired level with the increased sample size in many of the scenarios of unbiased estimate of treatment effect – especially in scenarios of MCAR, MAR with the same direction of dropout, and MNAR with the same direction of dropout and equal dropout rate. In scenarios of MAR with opposite direction of dropout, the increased sample for 30% dropout rate was not sufficient to reach the desired level of statistical power; however the observed power was not substantially lower than 85% when the desired power was 90% or 75% when the desired power was 80% in these scenarios

based on MMRM analysis. As noted earlier, CCA and MI yielded a slightly larger loss of power under these scenarios, and the power of MI can be improved by considering a substantially large number of imputations. Considering that the scenarios of opposite direction of dropout were extreme cases and bearing in mind the limitations of previously proposed approaches (Lu et al., 2008; 2009), which were detailed in chapter 2, the naïve approach of sample size inflation might be sufficient in order to account for expected attrition effect and provide an acceptable statistical power conditional on parameters that are less likely to adversely affect the probability of rejecting the null.

8.3.5 Reminder data to investigate the appropriateness of MAR-based analyses

From the simulation work, it was clear that the validity of approaches to handling missing data depend on the missing data mechanism. That is, to understand the potential impact of, and how best to deal with, missing data, it is important to consider the mechanism leading to the missingness. As pointed out in the background chapter, although some methods (Little, 1988; Diggle, 1989; Ridout, 1991; Fairclough, 2002) have been proposed for the identification of the missing data mechanism, their purpose is generally to detect violations of the MCAR assumption by identifying dependence on observed data. Fielding et al. (2009) performed an investigation of missing data mechanisms on a number of empirical trial datasets using these methods and found that different approaches yielded different conclusions on the mechanism among their datasets (detailed in chapter 7, discussion section). As there is no effective way to identify the underlying missing data mechanism with certainty and no best single analysis method under an MNAR mechanism, a primary analysis based on an MAR assumption is often reasonable (National Research Council, 2010; Mallinckrodt, 2013). Since the validity of inferences from this analysis depends on the correctness of the MAR assumption, it is critical to assess the robustness of inferences from the primary analysis to departures from MAR assumptions (Food and Drug Administration, 2008; European Medicines Agency, 2010;

National Research Council, 2010). However, due to lack of consensus on sensitivity analysis methods to assess the robustness of results from primary analysis of an RCT and how to draw inferences from these analyses, sensitivity analyses are often performed inappropriately and largely unreported (Morris et al., 2014). A simple approach has been proposed in chapter 7 to ensure the ignorability that is assumed with the primary analysis and thus to verify the unbiasedness of the estimate of treatment effect from the primary analysis if the possibility of bias from other (otherwise non-observable) sources has already been considered. A more detailed discussion was provided in chapter 7.

The proposed approach makes use of the responses subsequently recovered via reminder or MDC to assess the validity of the inferences from the MAR-based primary analysis. In the TATE trial, this approach showed that an MAR-based analysis was unlikely to yield unbiased estimates of treatment effect for the primary and secondary outcomes (i.e. non-responses were non-ignorable under the MAR-based analysis). Therefore, further sensitivity analyses were required under a range of plausible MNAR assumptions. However, in the STarT Back trial, this approach showed that an MAR-based analysis was unlikely to yield a biased estimate of treatment effect for the primary outcome (hence, non-responses were ignorable under MAR).

In summary, this analysis takes a position in between the primary analysis and a detailed sensitivity analysis, and allows a decision to be taken on whether to move ahead with the detailed sensitivity analysis that makes stringent assumptions about missing data. Guidance under the proposed approach is based on the assumption that reminder responses share the same (or at least similar) mechanism underlying missingness as the true non-responders/missing data.

8.4 Limitations and generalizability

The present systematic review was restricted to trials published in speciality journals relating to musculoskeletal/pain disorders. The selected journals were high-impact factor journals, which should bias the results towards a better methodology and reporting; hence, it is expected that the systematic review findings around appropriate use of methodology for ITT analysis and proper handling of missing data will be overstated rather than understated. Thus, the finding in this review that the minority of RCTs are performing a full ITT analysis (as per recommendation) is likely to be conservative, particularly in respect of publications within speciality journals in the area of musculoskeletal clinical trials. Further, the systematic review of trials published in 2010 and 2011 does not necessarily reflect the impact of recent regulatory guidelines on the prevention and handling of missing data (European Medicines Agency, 2010; National Research Council, 2010) and thus the quality of practice in 2014. However, the systematic review provides a strong indication that inappropriate practices in dealing with missing data still exist and that little progress has been made in reducing the large proportion of trials that are inappropriately analysed compared to the similar previous reviews in this area that evaluated trials published in 2001 and 2002 (Wood et al., 2004; Gravel et al., 2007). Furthermore, the present review can be the reference for a future systematic review that could assess the impact of the regulatory guidelines on the quality of practice with regard to the treatment of missing data.

Though the focus of the thesis is on musculoskeletal trials, the simulation study findings can be generalized to any parallel-group trials with serial multiple outcome measurements with a continuous outcome variable that is assumed to be normally distributed. The clinical trial scenarios considered are likely to reflect some of the most common scenarios encountered across the spectrum of real-life pragmatic clinical trials where data is collected at set intervals and missing data is commonplace. The simulation study does

have limitations and does not reflect the full range of possible scenarios in RCTs. The findings of the study are limited in interpretation to RCTs with normally distributed outcomes and they should not be generalized to RCTs in which the outcome is non-normal; for example, those that involve substantial skew, a binary outcome or in which time-to-event is the primary outcome. For selecting trajectories, the present simulation study did not consider a common scenario of a short-term treatment effect that disappears. However, the simulation study did show that the shape of trajectories does not influence bias in either the ANCOVA or the MMRM estimate of difference in treatment effects between groups at a given time point unless missing responses were imputed using LOCF. Further, this study may not reflect all reasons for missing data and thus all patterns of missing data that could occur in real situations. However, in order to overcome this limitation to some extent, the simulation study has considered contrasting scenarios (e.g. 10% versus 30% overall dropout rates, equal versus differential dropout rates between groups, same versus opposite direction of dropouts) in generating missing data. This study considered missing data due to dropouts but not due to missed visits or missing item(s). However, the statistical methods that have been considered in this study are generalizable to non-monotone missing data as well. Another limitation is that the study did not consider other methods, such as weighted generalized estimating equations (wGEE; Mallinckrodt, 2013), for handling of missing data. The wGEE approach involves weighting observations by the inverse probability of dropout, and the validity of this approach depends on the correct specification of the dropout model (Mallinckrodt, 2013).

The illustration of the proposed approach to consideration of reminder data has only been performed on two musculoskeletal related, parallel-group trials where outcome measurements were assumed to be normally distributed and estimate was the difference in treatment effects between groups at a given time point. The effect sizes in these trials were also particularly small, which may limit the interpretation of difference in effect size between the original and the modified datasets. Therefore, this approach needs to be

evaluated in other trial contexts, including trials with different effect sizes and other types of design, and trials in other clinical areas. Further, the proposed approach that was used to confirm the ignorability of missing data mechanism using responses obtained after a number of reminders depends heavily on the assumption that the reminder responses are representatives of missing responses. In fact this assumption cannot be verified, as the data required to do so are missing. Collection and comparison of reasons for delayed responses and non-responses may help inform the justification for this assumption. Such a comparison was not performed in the present study because the reasons for delayed response and largely for non-response were not available for the TATE and STarT Back trials.

8.5 Implications for practice

The simulation study findings can act as a guide to the selection of the most appropriate statistical approach for dealing with missing data in clinical trials. Further, this study could address the conflicting findings from the previous studies. Importantly, this study found that equal dropout rate guarantees an unbiased estimate of treatment effect from CCA, MMRM or MI only for scenarios of the same direction of dropout in both treatment groups. Contrary to the findings from many of the previous studies, the present simulation study shows the non-conservative nature (i.e. bias in estimate favours the new treatment) of the LOCF imputation approach. This study further showed that MMRM and MI yield similar results if they are implemented appropriately. This finding indicates that one should not devalue MI on the basis of a finding by Siddiqui (2011) in which the author reported that MI was severely impaired by a very low type 1 error rate (i.e. MI was too conservative to reject a true null hypothesis). The previous study finding was highlighted as the drawback of MI over MMRM in a recent report (Gewandter et al., 2014); however, the claim is unwarranted. The present study also favours an MMRM over MI with a restrictive imputation modelling strategy considering the ease with which MMRM can be

implemented – a substantially high number of imputations are required with the MI to obtain an estimate that is equivalent to MMRM. The advantage of an MI-inclusive modelling strategy is that an imputation model can include auxiliary variables that may predict missing values; especially for the secondary variables when MDC is limited to the primary outcome variable, the 'additional information' on the primary outcome variable may help to improve the estimation of treatment effect on the secondary variables. However, the evaluation of the empirical datasets suggests that MI with an inclusive imputation modelling strategy may be a reasonable consideration for a sensitivity analysis because of the difficulty associated with the selection of the 'best' model to impute missing data. Further, this study encourages the use of the proposed reminder approach to check the ignorability that is assumed with MMRM analysis and thus to confirm the unbiasedness of the estimate of treatment effect from the MMRM analysis.

Given the findings of this study, the following are recommended for future statistical analysis of a normally-distributed incomplete continuous outcome in a longitudinal randomized clinical trial:

i. Minimize the amount of missing data: the best way to deal with missing data is to avoid them. Trialists should anticipate potential missing data problems. The study protocol should address this issue and show necessary steps taken in the design and conduct of the trial to limit the impact of missing data. In pragmatic trials, it is often possible to use reminders to encourage participants to respond.

ii. Perform a priori sample size calculation and ensure it is adjusted for anticipated loss to follow-ups. In this thesis, it has been showed that the naïve approach of inflating the sample size directly proportional to anticipated loss to follow-up is generally acceptable to retain sufficient power to detect the true difference if the missing data analysis method is appropriate.

iii. All randomized participants should be accounted for in reporting the trial results irrespective of whether or not the participants are lost to follow-up. In the systematic review, it was found that the use of partial intention-to-treat analysis is quite common. Any planned exclusion of randomized participants from the analysis should be pre-specified in the study protocol with justification for such exclusions.

iv. Statistical methods to deal with missing data – the primary analysis and sensitivity analyses – should be pre-specified in the study protocol, and the assumptions required for such analyses should also be stated clearly.

v. LOCF-based analysis should not be regarded as the primary analysis. The analysis provides biased estimates with underestimated SEs in most situations; the estimate of treatment effect is not conservative in many situations.

vi. CCA should also not be performed as the primary analysis because MAR-based analyses – such as MMRM and MI-based analysis – can efficiently handle missing data more appropriately and in a greater number of situations than CCA, and retain sufficient CI coverage and statistical power. Missing data have been shown in many cases to at least follow an MAR mechanism, and therefore CCA is inadequate for analysing such datasets.

vii. Equal dropout rate between treatment groups does not guarantee an unbiased estimate of treatment effect (notably as dropout may not be in the same direction in both treatment groups).

viii. An MAR-based analysis might be ideal as the primary analysis because MNAR-based analyses require stringent assumptions about missing data. Unless a researcher is more comfortable with MI than with mixed models or covariates having missing observations, it is better to adopt an MMRM model

as the primary analysis. However, MI with an inclusive modelling strategy can be performed as a sensitivity analysis.

ix. It is advisable to perform the proposed approach (as outlined in chapter 7), which uses reminder responses to check the ignorability that is assumed with MMRM analysis and thus to confirm, or not, the unbiasedness of the estimate of treatment effect from the MMRM analysis.

x. If the proposed approach does not favour the ignorability of the missing data mechanism and the unbiasedness of the estimate of treatment effect, well-structured sensitivity analyses (Mallinckrodt et al., 2013; Ratitch et al., 2013) that include statistical models that incorporate plausible departure from the ignorability of the missing data mechanism should be included as part of the statistical analysis plan.

8.6 Future work

Further additional research is required to address the limitations of the present work.

i. A further systematic review is required, considering the present review as a reference, to assess the implication of recent research and regulatory guidelines with regard to missing data in clinical trials.

ii. The present simulation study did not consider the availability of multiple outcome measures, covariates and auxiliary variables. Therefore, this study did not explore the advantages of MI with inclusive imputation modelling strategies. Since the validity of MI involves the correctness of the imputation model, guidelines on selection of auxiliary variables into an imputation model are required.

iii. The rationale underlying the proposed approach is that the responses that have been retrieved after a number of failed attempts are likely to represent the

actual missing responses when the number of failed attempts increases. It would be interesting to investigate the plausibility of this assumption by utilizing reasons for delayed responses and non-responses.

iv. If the proposed approach does not support the ignorability of the missing data mechanism and the unbiasedness of the estimate of treatment effect, further investigation is needed on how to make best use of the reminder responses within modelling strategies that appropriately take into account departure from the ignorability of the missing data mechanism.

8.7 Conclusion

Given that no method can completely overcome the problem of missing data, trialists should consider ways to prevent missing data during the design and conduct of RCTs. In a pragmatic setting, it is quite often possible to consider sending reminders to encourage participants to respond and, as a final attempt, to limit the data collection to essential variables of interest, in order to maximize response to the primary endpoint.

A Monte Carlo simulation study was conducted to compare four methods for dealing with missing longitudinal normal continuous outcome data in clinical trials. On the basis of this, LOCF ANCOVA is very unlikely to give an unbiased estimate of treatment effect. Differential improvement in outcome between treatment groups and timing of dropout substantially affect the estimation of treatment effect using LOCF-based analysis. Hence, the use of LOCF-based analysis as the primary analysis in an RCT should be avoided.

Although CCA can produce an unbiased estimate of treatment effect in many scenarios, the analysis does not consider the availability of outcome responses at interim visits. In many situations in clinical trials, these interim measurements have some level of influence on participants' decision to continue a trial. Therefore, CCA should be disregarded in favour of more efficient analysis methods such as MMRM and MI-based analysis that can

effectively utilize the additional available data. MMRM and MI with a restrictive imputation modelling strategy yield similar results when they are implemented in a similar manner. In an MMRM model, baseline can be considered either as covariate or as outcome. Both strategies are acceptable if baseline data are complete, otherwise a model that includes baseline as an outcome is preferable.

The proposed reminder approach can be used to assess the robustness of the MAR assumption by checking expected consistency in MMRM estimates. If the results deviate, then analyses incorporating a range of plausible MNAR assumptions are advisable, at least as sensitivity tests for the evaluation of treatment effect.

References

Abraha, L. & Montedori, A. (2010). Modified intention to treat reporting in randomised controlled trials: systematic review. *BMJ.* 340: c2697.

Akai, M. (2010). Musculoskeletal Disorders. *Pharmaceutical Sciences Encyclopedia.* 224: 1-23.

Akl, E., Briel, M., You, J., Lamontagne, F., Gangji, A., Cukierman-Yaffe, T., Alshurafa, M., Sun, X., Nerenberg, K., Johnston, B., Vera, C., Mills, E., Bassler, D., Salazar, A., Bhatnagar, N., Busse, J., Khalid, Z., Walter, S.D., Cook, D., Schunemann, H., Altman, D. & Guyatt, G. (2009). Lost to follow-up information in trials (LOST-IT): a protocol on the potential impact. *Trials.* 10: 40.

Allison, P.D. (2001). *Missing data.* Thousand Oaks, CA: Sage.

Alshurafa, M., Briel, M., Akl, E.A., Haines, T., Moayyedi, P., Gentles, S.J., Rios, L., Tran, C., Bhatnagar, N., Lamontagne, F., Walter, S.D. & Guyatt, G.H. (2012). Inconsistent definitions for intention-to-treat in relation to missing outcome data: systematic review of the methods literature. *PLoS ONE.* 7: e49163.

Altman, D. (2009). Missing outcomes in randomized trials: addressing the dilemma. *Open Medicine.* 3: e51-53.

Baerwald, C., Verdecchia, P., Duquesroix, B., Frayssinet, H. & Ferreira, T. (2010). Efficacy, safety, and effects on blood pressure of naproxcinod 750 mg twice daily compared with placebo and naproxen 500 mg twice daily in patients with osteoarthritis of the hip: a randomized, double-blind, parallel-group, multicenter study. *Arthritis & Rheumatism.* 62: 3635-44.

Bala, M.M., Akl, E.A., Sun, X., Bassler, D., Mertz, D., Mejza, F., Vandvik, P.O., Malaga, G., Johnston, B.C., Dahm, P., Alonso-Coello, P., Diaz-Granados, N., Srinathan, S.K., Hassouneh, B., Briel, M., Busse, J.W., You, J.J., Walter, S.D., Altman, D.G. & Guyatt, G.H. (2013). Randomized trials published in higher vs. lower impact journals differ in design, conduct, and analysis. *Journal of Clinical Epidemiology.* 66: 286-95.

Barnes, S.A., Mallinckrodt, C.H., Lindborg, S.R. & Carter, M.K. (2008). The impact of missing data and how it is handled on the rate of false-positive results in drug development. *Pharmaceutical Statistics.* 7: 215-25.

Baron, G., Boutron, I., Giraudeau, B. & Ravaud, P. (2005). Violation of the intent-to-treat principle and rate of missing data in superiority trials assessing structural outcomes in rheumatic diseases. *Arthritis & Rheumatism.* 52: 1858-65.

Baron, G., Ravaud, P., Samson, A. & Giraudeau, B. (2008). Missing data in randomized controlled trials of rheumatoid arthritis with radiographic outcomes: a simulation study. *Arthritis Care & Research.* 59: 25-31.

Beaudreuil, J., Lasbleiz, S., Richette, P., Seguin, G., Rastel, C., Aout, M., Vicaut, E., Cohen-Solal, M., Lioté, F., de Vernejoul, M., Bardin, T. & Orcel, P. (2011). Assessment of dynamic humeral centering in shoulder pain with impingement syndrome: a randomised clinical trial. *Annals of the Rheumatic Diseases.* 70: 1613-8.

Bell, M.L., Kenward, M.G., Fairclough, D.L. & Horton, N.J. (2013). Differential dropout and bias in randomised controlled trials: when it matters and when it may not. *BMJ.* 346:e8668.

Beunckens, C., Molenberghs, G. & Kenward, M.G. (2005). Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials.* 2: 379-86.

Birhanu, T., Molenberghs, G., Sotto, C. & Kenward, M.G. (2011). Doubly robust and multiple-imputation-based generalized estimating equations. *Journal of Biopharmaceutical Statistics.* 21: 202-25.

Bliddal, H., Leeds, A.R., Stigsgaard, L., Astrup, A. & Christensen, R. (2011). Weight loss as treatment for knee osteoarthritis symptoms in obese patients: 1-year results from a randomised controlled trial. *Annals of the Rheumatic Diseases.* 70: 1798-803.

Bodner, T.E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal.* 15: 651-75.

Borm, G.F., Fransen, J. & Lemmens, W.A.J.G. (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. *Journal of Clinical Epidemiology.* 60: 1234-8.

Bredemeier, M. (2012). Last observation carried forward approach threatens the validity of intent-to-treat analysis in fibromyalgia trials: Comment on the article by Arnold et al. *Arthritis & Rheumatism.* 64: 2809-10.

Burton, A., Altman, D.G., Royston, P. & Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine.* 25: 4279-92.

Carpenter, J.R. & Kenward, M.G. (2013). *Multiple imputation and its application.* Chichester, UK: Wiley.

Chan, A., Tetzlaff, J.M., Gøtzsche, P.C., Altman, D.G., Mann, H., Berlin, J.A., Dickersin, K., Hróbjartsson, A., Schulz, K.F., Parulekar, W.R., Krleža-Jerić, K., Laupacis, A. & Moher, D. (2013). SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ.* 346: e7586.

Chesterton, L.S., Lewis, A.M., Sim, J., Mallen, C.D., Mason, E.E., Hay, E.M. & van der Windt, D.A. (2013). Transcutaneous electrical nerve stimulation as adjunct to primary care management for tennis elbow: pragmatic randomised controlled trial (TATE trial). *BMJ (Clinical Research ed.).* 347: f5160.

Chesterton, L.S., van der Windt, D.A., Sim, J., Lewis, M., Mallen, C.D., Mason, E.E., Warlow, C., Vohora, K. & Hay, E.M. (2009). Transcutaneous electrical nerve stimulation for the management of tennis elbow: a pragmatic randomized controlled trial: the TATE trial (ISRCTN 87141084). *BMC Musculoskeletal Disorders.* 10: 156.

Collins, L.M., Schafer, J.L. & Kam, C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods.* 6: 330-51.

Committee for Proprietary Medical Products (2001). Points to consider on missing data. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003641.pdf [Accessed: 20 May 2012].

Crutzen, R., Viechtbauer, W., Kotz, D. & Spigt, M. (2013). No differential attrition was found in randomized controlled trials published in general medical journals: a meta-analysis. *Journal of Clinical Epidemiology.* 66: 948-54.

Daniels, S.E., Goulder, M.A., Aspley, S. & Reader, S. (2011). A randomised, five-parallel-group, placebo-controlled trial comparing the efficacy and tolerability of analgesic combinations including a novel single-tablet combination of ibuprofen/paracetamol for postoperative dental pain. *Pain.* 152: 632-42.

Davis, S. (2014). Mixed models for repeated measures using categorical time effects (MMRM). In: O'Kelly, M. & Ratitch, B. (eds.) *Clinical trials with missing data: a guide for practitioners.* Chichester, UK: Wiley.

Deo, A., Schmid, C.H., Earley, A., Lau, J. & Uhlig, K. (2011). Loss to analysis in randomized controlled trials in CKD. *American Journal of Kidney Diseases.* 58: 349-55.

DeSouza, C.M., Legedza, A.T.R. & Sankoh, A.J. (2009). An overview of practical approaches for handling missing data in clinical trials. *Journal of Biopharmaceutical Statistics.* 19: 1055-73.

Diggle, P.J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics.* 45: 1255-8.

Dinh, P. & Yang, P. (2011). Handling baselines in repeated measures analyses with missing data at random. *Journal of Biopharmaceutical Statistics.* 21: 326-41.

Dwan, K., Altman, D.G., Arnaiz, J.A., Bloom, J., Chan, A.W., Cronin, E., Decullier, E., Easterbrook, P.J., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J.P., Simes, J. & Williamson, P.R. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS ONE.* 3: e3081.

Dziura, J.D., Post, L.A., Zhao, Q., Fu, Z. & Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale Journal of Biology and Medicine.* 86: 343-58.

Egbewale, B.E., 2012. *Statistical analysis of randomized controlled trials: a simulation and empirical study of methods of covariate adjustment*. PhD Thesis, Keele University.

Egbewale, B.E., Lewis, M. & Sim, J. (2014). Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. *BMC Medical Research Methodology.* 14: 49.

Enders, C.K. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

Enders, C.K. & Gottschall, A.C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling–a Multidisciplinary Journal.* 18: 35-54.

European Medicines Agency (2010). Guideline on missing data in confirmatory clinical trials. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf [Accessed: 15 April 2012].

European Medicines Agency (2006). ICH topic E9: statistical principles for clinical trials. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf [Accessed: 15 April 2012].

Fairclough, D.L. (2002). *Design and analysis of quality of life studies in clinical trials.* Boca Raton, FL: Chapman & Hall/CRC.

Fary, R.E., Carroll, G.J., Briffa, T.G. & Briffa, N.K. (2011). The effectiveness of pulsed electrical stimulation in the management of osteoarthritis of the knee: results of a double-blind, randomized, placebo-controlled, repeated-measures trial. *Arthritis & Rheumatism.* 63: 1333-42.

Feinman, R.D. (2009). Intention-to-treat. What is the question? *Nutrition & Metabolism.* 6: 1.

Fielding, S., Fayers, P. & Ramsay, C. (2010). Predicting missing quality of life data that were later recovered: an empirical comparison of approaches. *Clinical Trials.* 7: 333-42.

Fielding, S., Fayers, P.M. & Ramsay, C.R. (2009). Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health and Quality of Life Outcomes.* 7: 57.

Fielding, S., Maclennan, G., Cook, J.A. & Ramsay, C.R. (2008). A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials.* 9: 51.

Fielding, S., Fayers, P. & Ramsay, C.R. (2012). Analysing randomised controlled trials with missing data: choice of approach affects conclusions. *Contemporary Clinical Trials.* 33: 461-9.

Fisher, L.D., Dixon, D.O., Herson, J., Frankowski, R.K., Hearon, M.S. & Pearce, K.E. (1990). Intention to treat in clinical trials. In: Pearce, K.E. (ed.) *Statistical issues in drug research and development.* New York, NY: Marcel Dekker.

Fitzmaurice, G., Laird, N. & Ware, J. (2004). *Applied longitudinal analysis.* Hoboken, NJ: Wiley.

Fleming, T.R. (2011). Addressing missing data in clinical trials. *Annals of Internal Medicine.* 154: 113-7.

Food and Drug Administration (2008). Guidance for sponsors, clinical investigators, and IRBs: data retention when subjects withdraw from FDA-regulated clinical trials. Available at: http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm126489.pdf [Accessed: 10 May 2012].

Food and Drug Administration (1997). International conference on harmonisation: guidelines on general considerations for clinical trials. *Federal Register.* 62: 66113-9.

Frangakis, C. & Rubin, D. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika.* 86: 365-79.

Fritsche, G., Frettlöh, J., Hüppe, M., Dlugaj, M., Matatko, N., Gaul, C. & Diener, H. (2010). Prevention of medication overuse in patients with migraine. *Pain.* 151: 404-13.

Gabriel, A.P. & Mercado, C.P. (2011). Data retention after a patient withdraws consent in clinical trials. *Open Access Journal of Clinical Trials.* 3: 15-9.

Genevay, S., Viatte, S., Finckh, A., Zufferey, P., Balagué, F. & Gabay, C. (2010). Adalimumab in severe and acute sciatica: a multicenter, randomized, double-blind, placebo-controlled trial. *Arthritis & Rheumatism.* 62: 2339-46.

Gewandter, J.S., McDermott, M.P., McKeown, A., Smith, S.M., Williams, M.R., Hunsinger, M., Farrar, J., Turk, D.C. & Dworkin, R.H. (2014). Reporting of missing data and methods used to accommodate them in recent analgesic clinical trials: ACTTION systematic review and recommendations. *Pain.* 155: 1871-7.

Graham, J.W., Olchowski, A.E. & Gilreath, T.D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science.* 8: 206-13.

Gravel, J., Opatrny, L. & Shapiro, S. (2007). The intention-to-treat approach in randomized controlled trials: are authors saying what they do and doing what they say? *Clinical Trials.* 4: 350-6.

Hardt, J., Herke, M. & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Medical Research Methodology.* 12: 184.

Hay, E.M., Dunn, K.M., Hill, J.C., Lewis, M., Mason, E.E., Konstantinou, K., Sowden, G., Somerville, S., Vohora, K., Whitehurst, D. & Main, C.J. (2008). A randomised clinical trial of subgrouping and targeted treatment for low back pain compared with best current care. The STarT Back trial study protocol. *BMC Musculoskeletal Disorders.* 9: 58.

Hedeker, D. & Gibbons, R.D. (2006). *Longitudinal data analysis.* Hoboken, NJ: Wiley.

Henschke, N., Kuijpers, T., Rubinstein, S.M., van Middelkoop, M., Ostelo, R., Verhagen, A., Koes, B.W. & van Tulder, M.W. (2012). Trends over time in the size and quality of randomised controlled trials of interventions for chronic low-back pain. *European Spine Journal.* 21: 375-81.

Heritier, S.R., Gebski, V.J. & Keech, A.C. (2003). Inclusion of patients in clinical trial analysis: the intention-to-treat principle. *Medical Journal of Australia.* 179: 438-40.

Hewlett, S., Ambler, N., Almeida, C., Cliss, A., Hammond, A., Kitchen, K., Knops, B., Pope, D., Spears, M., Swinkels, A. & Pollock, J. (2011). Self-management of fatigue in rheumatoid arthritis: a randomised controlled trial of group cognitive-behavioural therapy. *Annals of the Rheumatic Diseases.* 70: 1060-7.

Higgins, J.P.T. & Altman, D.G. (2011). Chapter 8: Assessing risk of bias in included studies. In: Higgins, J.P.T. & Green, S. (eds.) *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0.* The Cochrane Collaboration. Accessed at: www.cochrane-handbook.org [Accessed: 08 June 2014].

Hill, J.C., Whitehurst, D.G., Lewis, M., Bryan, S., Dunn, K.M., Foster, N.E., Konstantinou, K., Main, C.J., Mason, E., Somerville, S., Sowden, G., Vohora, K. & Hay, E.M. (2011). Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet.* 378: 1560-71.

Hollis, S. & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ.* 319: 670.

Hopewell, S., Collins, G., Hirst, A., Kirtley, S., Tajar, A., Gerry, S. & Altman, D. (2013). Reporting characteristics of non-primary publications of results of randomized trials: a cross-sectional review. *Trials.* 14: 240.

Hopewell, S., Hirst, A., Collins, G., Mallett, S., Yu, L. & Altman, D. (2011). Reporting of participant flow diagrams in published reports of randomized trials. *Trials.* 12: 253.

Horton, N.J. & Lipsitz, S.R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing data. *American Statistician.* 55: 244-54.

International Conference on Harmonisation (1998). Statistical Principles for Clinical trials, recommended for adoption to the regulatory bodies of the European Union, Japan and USA. Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf [Accessed: 12 July 2012].

Ioannidis, J.P., Evans, S.J., Gotzsche, P.C., O'Neill, R.T., Altman, D.G., Schulz, K., Moher, D. & CONSORT Group (2004). Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Annals of Internal Medicine.* 141: 781-8.

Johnsen, A.T., Petersen, M.A., Gluud, C., Lindschou, J., Fayers, P., Sjogren, P., Pedersen, L., Neergaard, M.A., Vejlgaard, T.B., Damkier, A., Nielsen, J.B., Stromgren, A.S., Higginson, I.J. & Groenvold, M. (2014). Detailed statistical analysis plan for the Danish Palliative Care Trial (DanPaCT). *Trials.* 15: 376.

Kenward, M.G. & Roger, J.H. (1997). Small sample Inference for fixed effects from restricted maximum likelihood. *Biometrics.* 53: 983-97.

Kenward, M.G. & Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research.* 16: 199-218.

Kenward, M.G., White, I.R. & Carpenter, J.R. (2010). Letter to the editor: Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? by G. F. Liu, K. Lu, R. Mogg, M. Mallick and D. V. Mehrotra, Statistics in Medicine 2009; 28:2509-2530. *Statistics in Medicine.* 29: 1455-6.

Kim, M. (2006). Statistical methods in Arthritis and Rheumatism: current trends. *Arthritis & Rheumatism.* 54: 3741-9.

Kim, Y. (2011). Missing data handling in chronic pain trials. *Journal of Biopharmaceutical Statistics.* 21: 311-25.

Koes, B.W., Malmivaara, A. & van Tulder, M.W. (2005). Trend in methodological quality of randomised clinical trials in low back pain. *Best Practice & Research Clinical Rheumatology.* 19: 529-39.

Kruse, R.L., Alper, B.S., Reust, C., Stevermer, J.J., Shannon, S. & Williams, R.H. (2002). Intention-to-treat analysis: Who is in? Who is out? *Journal of Family Practice.* 51: 969-71.

Lachin, J.M. (2000). Statistical consideration in the intention-to-treat principle. *Controlled Clinical Trials.* 21: 167-89.

Laird, N.M. & Ware, H.J. (1982). Random-effects models for longitudinal data. *Biometrics.* 38: 963-74.

Lane, P. (2008). Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. *Pharmaceutical Statistics.* 7: 93-106.

Lavori, P.W., Brown, C.H., Duan, N., Gibbons, R.D. & Greenhouse, J. (2008). Missing data in longitudinal clinical trials part A: design and conceptual issues. *Psychiatric annals.* 38: 784-92.

Lee, K.P., Schotland, M., Bacchetti, P. & Bero, L.A. (2002). Association of journal quality indicators with methodological quality of clinical research articles. *JAMA.* 287: 2805-8.

Lee, K.J. & Carlin, J.B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology.* 171: 624-32.

Lewis, J.A. & Machin, D. (1993). Intention to treat–who should use ITT? *British Journal of Cancer.* 68: 647-50.

Li, T., Hutfless, S., Scharfstein, D.O., Daniels, M.J., Hogan, J.W., Little, R.J.A., Roy, J.A., Law, A.H. & Dickersin, K. (2014). Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus. *Journal of Clinical Epidemiology.* 67: 15-32.

Liang, K.Y. & Zeger, S. (2000). Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhya: The Indian Journal of Statistics, Series B.* 62: 134-48.

Liang, K. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika.* 73: 13-22.

Little, R.J.A. (1995). Modelling the drop-out mechanism in repeated measures studies. *Journal of American Statistical Association.* 90: 1112-21.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data*. Chichester, UK: Wiley.

Little, R.J., D'Agostino, R., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., Frangakis, C., Hogan, J.W., Molenberghs, G., Murphy, S.A., Neaton, J.D., Rotnitzky, A., Scharfstein, D., Shih, W.J., Siegel, J.P. & Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine.* 367: 1355-60.

Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of American Statistical Association.* 83: 1198-202.

Little, R.J.A. & Rubin, D.B. (2002). *Statistical analysis with missing data.* 2nd edn. Hoboken, N.J: Wiley.

Liu, G.F., Lu, K., Mogg, R., Mallick, M. & Mehrotra, D.V. (2009). Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? *Statistics in Medicine.* 28: 2509-30.

Lu, K., Luo, X. & Chen, P. (2008). Sample size estimation for repeated measures analysis in randomized clinical trials with missing data. *International Journal of Biostatistics.* 4: 9.

Lu, K., Mehrotra, D.V. & Liu, G. (2009). Sample size determination for constrained longitudinal data analysis. *Statistics in Medicine.* 28: 679-99.

Lyass, A., 2010. *Assessing if randomized treatment group should be included in the imputation model when imputing missing outcome data in randomized superiority clinical trials.* PhD Thesis, Boston University.

Mackinnon, A. (2010). The use and reporting of multiple imputation in medical research – a review. *Journal of Internal Medicine.* 268: 586-93.

Mallinckrodt, C.H. (2013). *Preventing and treating missing data in longitudinal clinical trials: A practical guide.* Cambridge, UK: Cambridge University Press.

Mallinckrodt, C.H., Lin, Q. & Molenberghs, M. (2013). A structured framework for assessing sensitivity to missing data assumptions in longitudinal clinical trials. *Pharmaceutical Statistics.* 12: 1-6.

Mallinckrodt, C.H., Clark, W.S. & David, S.R. (2001a). Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics.* 11: 9-21.

Mallinckrodt, C.H., Clark, W.S. & David, S.R. (2001b). Type I error rates from mixed effects model repeated measures versus fixed effects ANOVA with missing values imputed via last observation carried forward. *Drug Information Journal.* 35: 1215-25.

Mallinckrodt, C.H., Kaiser, C.J., Watkin, J.G., Molenberghs, G. & Carroll, R.J. (2004). The effect of correlation structure on treatment contrasts estimated from incomplete clinical trial data with likelihood-based repeated measures compared with last observation carried forward ANOVA. *Clinical Trials.* 1: 477-89.

Mallinckrodt, C.H., Lane, P.W., Schnell, D., Peng, Y. & Mancuso, J.P. (2008). Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information Journal.* 42: 303-19.

Mallinckrodt, C.H., Sanger, T.M., Dubé, S., DeBrota, D.J., Molenberghs, G., Carroll, R.J., Potter, W.Z. & Tollefson, G.D. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry.* 53: 754-60.

McNamee, R. (2009). Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. *Statistics in Medicine.* 28: 2639-52.

Moher, D., Schulz, K.F. & Altman, D.G. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet.* 357: 1191-4.

Moher, D., Hopewell, S., Kenneth, F.S., Montori, V., Peter, C.G., Devereaux, P.J., Elbourne, D., Egger, M. & Douglas, G.A. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ.* 340: c869.

Molenberghs, G. & Kenward, M.G. (2007). *Missing data in clinical studies.* Chichester, UK: Wiley.

Montori, V.M. & Guyatt, G.H. (2001). Intention-to-treat principle. *Canadian Medical Association Journal.* 165: 1339-41.

Moore, R.A., Derry, S. & McQuay, H.J. (2008). Discontinuation rates in clinical trials in musculoskeletal pain: meta-analysis from etoricoxib clinical trial reports. *Arthritis Research & Therapy.* 10: R53.

Morris, T.P., Kahan, B.C. & White, I.R. (2014). Choosing sensitivity analyses for randomised trials: principles. *BMC Medical Research Methodology.* 14: 11.

National Research Council (2010). *The prevention and treatment of missing data in clinical trials. panel on handling missing data in clinical trials. Committee on national statistics, division of behavioral and social sciences and education.* Washington, DC: National Academies Press.

Navarro-Sarabia, F., Coronel, P., Collantes, E., Navarro, F.J., de la Serna, A.R., Naranjo, A., Gimeno, M. & Herrero-Beaumont, G. (2011). A 40-month multicentre, randomised placebo-controlled study to assess the efficacy and carry-over effect of repeated intra-articular injections of hyaluronic acid in knee osteoarthritis: the AMELIA project. *Annals of the Rheumatic Diseases.* 70: 1957-62.

Olsen, I.C., Kvien, T.K. & Uhlig, T. (2012). Consequences of handling missing data for treatment response in osteoarthritis: a simulation study. *Osteoarthritis and Cartilage.* 20: 822-8.

Peters, S.A., Bots, M.L., den Ruijter, H.M., Palmer, M.K., Grobbee, D.E., Crouse, J.R.,3rd, O'Leary, D.H., Evans, G.W., Raichlen, J.S., Moons, K.G., Koffijberg, H. & METEOR study group (2012). Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models. *Journal of Clinical Epidemiology.* 65: 686-95.

Prakash, A., Risser, R.C. & Mallinckrodt, C.H. (2008). The impact of analytic method on interpretation of outcomes in longitudinal clinical trials. *International Journal of Clinical Practice.* 62: 1147-58.

Ratitch, B. (2014). Multiple imputation. In: O'Kelly, M. & Ratitch, B. (eds.) *Clinical trials with missing data: a guide for practitioners.* Chichester, UK: Wiley.

Ratitch, B., O'Kelly, M. & Tosiello, R. (2013). Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics.* 12:337-47.

Ridout, M.S. (1991). Testing for random dropouts in repeated measurement data. *Biometrics.* 47: 1617-621.

Rubbert-Roth, A., Tak, P.P., Zerbini, C., Tremblay, J., Carreño, L., Armstrong, G., Collinson, N. & Shaw, T.M. (2010). Efficacy and safety of various repeat treatment dosing regimens of rituximab in patients with active rheumatoid arthritis: results of a Phase III randomized study (MIRROR). *Rheumatology.* 49: 1683-93.

Rubin, D. (1996). Multiple imputation after 18 years. *Journal of American Statistical Association.* 91: 473-90.

Rubin, D.B. (1987). *Multiple imputation for non-response in surveys.* New York, NY: Wiley.

Rubin, D.B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, pp. 20-8.

Rubin, D.B. (1976). Inference and missing data. *Biometrika.* 63: 581-92.

Russell, I.J., Holman, A.J., Swick, T.J., Alvarez-Horine, S., Wang, Y.G. & Guinta, D. (2011). Sodium oxybate reduces pain, fatigue, and sleep disturbance and improves functionality in fibromyalgia: results from a 14-week, randomized, double-blind, placebo-controlled study. *Pain.* 152: 1007-17.

SAS Institute Inc. (2011). *SAS 9.3.* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2014). *SAS/STAT 13.2 user's guide.* Cary, NC: SAS Institute Inc.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Schafer, J.L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research.* 8: 3-15.

Schafer, J.L. & Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological Methods.* 7: 147-77.

Schafer, J.L. & Olsen, M.K. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research.* 33: 545-71.

Schulz, K.F., Altman, D.G. & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ.* 340: c332.

Schulz, K.F., Grimes, D.A., Altman, D.G. & Hayes, R.J. (1996). Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *BMJ.* 312: 742-4.

Schwartz, D. & Lellouch, J. (2009). Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Clinical Epidemiology.* 62: 499-505.

Siddiqui, O. (2011). MMRM versus MI in dealing with missing data–a comparison based on 25 NDA data sets. *Journal of Biopharmaceutical Statistics.* 21: 423-36.

Siddiqui, O., Hung, H.M.J. & O'Neill, R. (2009). MMRM vs. LOCF: a comprehensive comparison based on simulation study and 25 NDA datasets. *Journal of Biopharmaceutical Statistics.* 19: 227-46.

Spratt, M., Carpenter, J., Sterne, J.A.C., Carlin, J.B., Heron, J., Henderson, J. & Tilling, K. (2010). Strategies for multiple imputation in longitudinal studies. *American Journal of Epidemiology.* 172: 478-87.

StataCorp (2013). *Stata 13 multiple-imputation reference manual.* College Station, TX: Stata Press.

StataCorp (2011). *Stata statistical software: Release 12.0.* College Station, TX: StataCorp LP.

Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M. & Carpenter, J.R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ.* 338: b2393.

Streiner, D.L. (2008). Missing data and the trouble with LOCF. *Evidence Based Mental Health.* 11: 3-5.

Thoemmes, F. & Rose, N. (2014). A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research.* 49: 443-59.

Thomson Reuters (2011). *Journal citation reports. Available at:* http:// thomsonreuters.com/products_services/science/science_products/a-z/journal_citation_reports [Accessed: 10 December 2011].

Thorn, B.E., Day, M.A., Burns, J., Kuhajda, M.C., Gaskins, S.W., Sweeney, K., McConley, R., Ward, L.C. & Cabbil, C. (2011). Randomized trial of group cognitive behavioral therapy compared with a pain education control for low-literacy rural people with chronic pain. *Pain.* 152: 2710-20.

Turner, L., Shamseer, L., Altman, D., Schulz, K. & Moher, D. (2012). Does use of the CONSORT statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Systematic Reviews.* 1: 60.

Twisk, J., de Boer, M., de Vente, W. & Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of Clinical Epidemiology.* 66: 1022-8.

Van Breukelen, G.J.P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology.* 59: 920-5.

Verbeke, G. & Molenberghs, G. (2005). Longitudinal and incomplete clinical studies. *International Journal of Statistics.* 63: 143-76.

Verstappen, S.M.M., McCoy, M.J., Roberts, C., Dale, N.E., Hassell, A.B. & Symmons, D.P.M. (2010). Beneficial effects of a 3-week course of intramuscular glucocorticoid injections in patients with very early inflammatory polyarthritis: results of the STIVEA trial. *Annals of the Rheumatic Diseases.* 69: 503-9.

Wertz, R.T. (1993). Intention to treat: once randomized, always analyzed. *Clinical Aphasiology.* 23: 57-64.

White, I.R., Carpenter, J. & Horton, N.J. (2012). Including all individuals is not enough: lessons for intention-to-treat analysis. *Clinical Trials.* 9: 396-407.

White, I.R., Horton, N.J., Carpenter, J. & Pocock, S.J. (2011). Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ.* 342: d40.

White, I.R., Royston, P. & Wood, A.M. (2011b). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine.* 30: 377-99.

White, I.R. & Thompson, S.G. (2005). Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine.* 24: 993-1007.

Wong, W.K., Boscardin, W.J., Postlethwaite, A.E. & Furst, D.E. (2011). Handling missing data issues in clinical trials for rheumatic diseases. *Contemporary Clinical Trials.* 32: 1-9.

Wood, A.M., White, I.R. & Thompson, S.G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials.* 1: 368-76.

Woolley, S.B., Cardoni, A.A. & Goethe, J.W. (2009). Last-observation-carried-forward imputation method in clinical efficacy trials: review of 352 antidepressant studies. *Pharmacotherapy.* 29: 1408-16.

Wright, C.C. & Sim, J. (2003). Intention-to-treat approach to data from randomized controlled trials: a sensitivity analysis. *Journal of Clinical Epidemiology.* 56: 833-42.

Appendices

Appendix 1: List of 91 reviewed trial publications 273

Appendix 2: Systematic review - data extraction form 281

Appendix 3: Stata syntax for data simulation and analyses 285

Appendix 4: Simulation results (tables 1 and 2) 310

Appendix 5: Simulation results (tables 3–22) 312

Appendix 6: MMRM versus cLDA (tables 23 and 24) 332

Appendix 7: Effect of sample size on statistical power (tables 25 and 26; figures 1 and 2) 333

Appendix 8: TATE trial: MI-inclusive imputation models (tables 27–30) 338

Appendix 9: TATE results (MMRM analysis of pain intensity score; table 31) 342

Appendix 1: List of 91 reviewed trial publications

Alavi, A., Goodfellow, L., Fraser, O., Tarelli, E., Bland, M. & Axford, J. (2011). A double-blind, randomized, placebo-controlled study to explore the efficacy of a dietary plant-derived polysaccharide supplement in patients with rheumatoid arthritis. *Rheumatology.* 50 (6): 1111-1119.

Alten, R.E., Zerbini, C., Jeka, S., Irazoque, F., Khatib, F., Emery, P., Bertasso, A., Rabbia, M. & Caulfield, J.P. (2010). Efficacy and safety of pamapimod in patients with active rheumatoid arthritis receiving stable methotrexate therapy. *Annals of the Rheumatic Diseases.* 69 (2): 364-367.

Andersen, L.L., Saervoll, C.A., Mortensen, O.S., Poulsen, O.M., Hannerz, H. & Zebis, M.K. (2011). Effectiveness of small daily amounts of progressive resistance training for frequent neck/shoulder pain: Randomised controlled trial. *Pain.* 152 (2): 440-446.

Arnold, L.M., Gendreau, R.M., Palmer, R.H., Gendreau, J.F. & Wang, Y. (2010). Efficacy and safety of milnacipran 100 mg/day in patients with fibromyalgia: Results of a randomized, double-blind, placebo-controlled trial. *Arthritis & Rheumatism.* 62 (9): 2745-2756.

Atchia, I., Kane, D., Reed, M.R., Isaacs, J.D. & Birrell, F. (2011). Efficacy of a single ultrasound-guided injection for the treatment of hip osteoarthritis. *Annals of the Rheumatic Diseases.* 70 (1): 110-116.

Baerwald, C., Verdecchia, P., Duquesroix, B., Frayssinet, H. & Ferreira, T. (2010). Efficacy, safety, and effects on blood pressure of naproxcinod 750 mg twice daily compared with placebo and naproxen 500 mg twice daily in patients with osteoarthritis of the hip: a randomized, double-blind, parallel-group, multicenter study. *Arthritis & Rheumatism.* 62 (12): 3635-3644.

Baranauskaite, A., Raffayová, H., Kungurov, N., Kubanova, A., Venalis, A., Helmle, L., Srinivasan, S., Nasonov, E. & Vastesaeger, N. (2011). Infliximab plus methotrexate is superior to methotrexate alone in the treatment of psoriatic arthritis in methotrexate-naive patients: the RESPOND study. *Annals of the Rheumatic Diseases.*

Beaudreuil, J., Lasbleiz, S., Richette, P., Seguin, G., Rastel, C., Aout, M., Vicaut, E., Cohen-Solal, M., Lioté, F., de Vernejoul, M., Bardin, T. & Orcel, P. (2011). Assessment of dynamic humeral centering in shoulder pain with impingement syndrome: a randomised clinical trial. *Annals of the Rheumatic Diseases.* 70 (9): 1613-1618.

Bingham, C.O., Looney, R.J., Deodhar, A., Halsey, N., Greenwald, M., Codding, C., Trzaskoma, B., Martin, F., Agarwal, S. & Kelman, A. (2010). Immunization responses in rheumatoid arthritis patients treated with rituximab: Results from a controlled clinical trial. *Arthritis & Rheumatism.* 62 (1): 64-74.

Bliddal, H., Leeds, A.R., Stigsgaard, L., Astrup, A. & Christensen, R. (2011). Weight loss as treatment for knee osteoarthritis symptoms in obese patients: 1-year results from a randomised controlled trial. *Annals of the Rheumatic Diseases.* 70 (10): 1798-1803.

Bos, W.H., Dijkmans, B.A.C., Boers, M., van de Stadt, R.J. & van Schaardenburg, D. (2010). Effect of dexamethasone on autoantibody levels and arthritis development in patients with arthralgia: a randomised trial. *Annals of the Rheumatic Diseases.* 69 (3): 571-574.

Branco, J.C., Zachrisson, O., Perrot, S. & Mainguy, Y. (2010). A European multicenter randomized double-blind placebo-controlled monotherapy clinical trial of milnacipran in treatment of fibromyalgia. *The Journal of Rheumatology.* 37 (4): 851-859.

Braun, J., van der Horst-Bruinsma, I.E., Huang, F., Burgos-Vargas, R., Vlahos, B., Koenig, A.S. & Freundlich, B. (2011). Clinical efficacy and safety of etanercept versus sulfasalazine in patients

with ankylosing spondylitis: A randomized, double-blind trial. *Arthritis & Rheumatism.* 63 (6): 1543-1551.

Brien, S., Lachance, L., Prescott, P., McDermott, C. & Lewith, G. (2011). Homeopathy has clinical benefits in rheumatoid arthritis patients that are attributable to the consultation process but not the homeopathic remedy: a randomized controlled clinical trial. *Rheumatology.* 50 (6): 1070-1082.

Brotz, D., Maschke, E., Burkard, S., Engel, C., Manz, C., Ernemann, U., Wick, W. & Weller, M. (2010). Is there a role for benzodiazepines in the management of lumbar disc prolapse with acute sciatica? *Pain.* 149 (3): 470-475.

Chao, J., Wu, C., Sun, B., Hose, M.K., Quan, A., Hughes, T.H., Boyle, D. & Kalunian, K.C. (2010). Inflammatory characteristics on ultrasound predict poorer long-term response to intraarticular corticosteroid injections in knee osteoarthritis. *The Journal of Rheumatology.* 37 (3): 650-655.

Chevalier, X., Jerosch, J., Goupille, P., van Dijk, N., Luyten, F.P., Scott, D.L., Bailleul, F. & Pavelka, K. (2010). Single, intra-articular treatment with 6 ml hylan G-F 20 in patients with symptomatic primary osteoarthritis of the knee: a randomised, multicentre, double-blind, placebo controlled trial. *Annals of the Rheumatic Diseases.* 69 (01): 113-119.

Christiansen, S., Oettingen, G., Dahme, B. & Klinger, R. (2010). A short goal-pursuit intervention to improve physical capacity: A randomized clinical trial in chronic back pain patients. *Pain.* 149 (3): 444-452.

Cunnington, J., Marshall, N., Hide, G., Bracewell, C., Isaacs, J., Platt, P. & Kane, D. (2010). A randomized, double-blind, controlled study of ultrasound-guided corticosteroid injection into the joint of patients with inflammatory arthritis. *Arthritis & Rheumatism.* 62 (7): 1862-1869.

Daniels, S.E., Goulder, M.A., Aspley, S. & Reader, S. (2011). A randomised, five-parallel-group, placebo-controlled trial comparing the efficacy and tolerability of analgesic combinations including a novel single-tablet combination of ibuprofen/paracetamol for postoperative dental pain. *Pain.* 152 (3): 632-642.

Distler, O., Eich, W., Dokoupilova, E., Dvorak, Z., Fleck, M., Gaubitz, M., Hechler, M., Jansen, J., Krause, A., Bendszus, M., Pache, L., Reiter, R. & Müller-Ladner, U. (2010). Evaluation of the efficacy and safety of terguride in patients with fibromyalgia syndrome: Results of a twelve-week, multicenter, randomized, double-blind, placebo-controlled, parallel-group study. *Arthritis & Rheumatism.* 62 (1): 291-300.

Doherty, M., Hawkey, C., Goulder, M., Gibb, I., Hill, N., Aspley, S. & Reader, S. (2011). A randomised controlled trial of ibuprofen, paracetamol or a combination tablet of ibuprofen/paracetamol in community-derived people with knee pain. *Annals of the Rheumatic Diseases.* 70 (9): 1534-1541.

Dougados, M., Braun, J., Szanto, S., Combe, B., Elbaz, M., Geher, P., Thabut, G., Leblanc, V. & Logeart, I. (2011). Efficacy of etanercept on rheumatic signs and pulmonary function tests in advanced ankylosing spondylitis: results of a randomised double-blind placebo-controlled study (SPINE). *Annals of the Rheumatic Diseases.* 70 (5): 799-804.

Elander, J., Robinson, G. & Morris, J. (2011). Randomized trial of a DVD intervention to improve readiness to self-manage joint pain. *Pain.* 152 (10): 2333-2341.

Emery, P., Deodhar, A., Rigby, W.F., Isaacs, J.D., Combe, B., Racewicz, A.J., Latinis, K., Abud-Mendoza, C., Szczepański, L.J., Roschmann, R.A., Chen, A., Armstrong, G.K., Douglass, W. & Tyrrell, H. (2010). Efficacy and safety of different doses and retreatment of rituximab: a randomised, placebo-controlled trial in patients who are biological naïve with active rheumatoid

arthritis and an inadequate response to methotrexate (Study Evaluating Rituximab's Efficacy in MTX iNadequate rEsponders (SERENE)). *Annals of the Rheumatic Diseases.* 69 (9): 1629-1635.

Fary, R.E., Carroll, G.J., Briffa, T.G. & Briffa, N.K. (2011). The effectiveness of pulsed electrical stimulation in the management of osteoarthritis of the knee: results of a double-blind, randomized, placebo-controlled, repeated-measures trial. *Arthritis & Rheumatism.* 63 (5): 1333-1342.

Forestier, R., Desfour, H., Tessier, J., Françon, A., Foote, A.M., Genty, C., Rolland, C., Roques, C. & Bosson, J. (2010). Spa therapy in the treatment of knee osteoarthritis: a large randomised multicentre trial. *Annals of the Rheumatic Diseases.* 69 (4): 660-665.

Fritsche, G., Frettlöh, J., Hüppe, M., Dlugaj, M., Matatko, N., Gaul, C. & Diener, H. (2010). Prevention of medication overuse in patients with migraine. *Pain.* 151 (2): 404-413.

Furie, R., Petri, M., Zamani, O., Cervera, R., Wallace, D.J., Tegzová, D., Sanchez-Guerrero, J., Schwarting, A., Merrill, J.T., Chatham, W.W., Stohl, W., Ginzler, E.M., Hough, D.R., Zhong, Z.J., Freimuth, W., van Vollenhoven, R.F. & BLISS-76 Study Group (2011). A phase III, randomized, placebo-controlled study of belimumab, a monoclonal antibody that inhibits B lymphocyte stimulator, in patients with systemic lupus erythematosus. *Arthritis & Rheumatism.* 63 (12): 3918-3930.

Gabay, C., Medinger-Sadowski, C., Gascon, D., Kolo, F. & Finckh, A. (2011). Symptomatic effects of chondroitin 4 and chondroitin 6 sulfate on hand osteoarthritis: A randomized, double-blind, placebo-controlled clinical trial at a single center. *Arthritis & Rheumatism.* 63 (11): 3383-3391.

Genevay, S., Viatte, S., Finckh, A., Zufferey, P., Balagué, F. & Gabay, C. (2010). Adalimumab in severe and acute sciatica: a multicenter, randomized, double-blind, placebo-controlled trial. *Arthritis & Rheumatism.* 62 (8): 2339-2346.

Ginzler, E.M., Wofsy, D., Isenberg, D., Gordon, C., Lisk, L. & Dooley, M. (2010). Nonrenal disease activity following mycophenolate mofetil or intravenous cyclophosphamide as induction treatment for lupus nephritis: Findings in a multicenter, prospective, randomized, open-label, parallel-group clinical trial. *Arthritis & Rheumatism.* 62 (1): 211-221.

Goldman, R.H., Stason, W.B., Park, S.K., Kim, R., Mudgal, S., Davis, R.B. & Kaptchuk, T.J. (2010). Low-dose amitriptyline for treatment of persistent arm pain due to repetitive use. *Pain.* 149 (1): 117-123.

Gray, P., Kirby, J., Smith, M.T., Cabot, P.J., Williams, B., Doecke, J. & Cramond, T. (2011). Pregabalin in severe burn injury pain: A double-blind, randomised placebo-controlled trial. *Pain.* 152 (6): 1279-1288.

Griffiths, B., Emery, P., Ryan, V., Isenberg, D., Akil, M., Thompson, R., Maddison, P., Griffiths, I.D., Lorenzi, A., Miles, S., Situnayake, D., Teh, L.S., Plant, M., Hallengren, C., Nived, O., Sturfelt, G., Chakravarty, K., Tait, T. & Gordon, C. (2010). The BILAG multi-centre open randomized controlled trial comparing ciclosporin vs azathioprine in patients with severe SLE. *Rheumatology.* 49 (4): 723-732.

Hartkamp, A., Geenen, R., Godaert, G.L.R., Bijl, M., Bijlsma, J.W.J. & Derksen, R.H.W.M. (2010). Effects of dehydroepiandrosterone on fatigue and well-being in women with quiescent systemic lupus erythematosus: a randomised controlled trial. *Annals of the Rheumatic Diseases.* 69 (6): 1144-1147.

Herrick, A.L., van den Hoogen, F., Gabrielli, A., Tamimi, N., Reid, C., O'Connell, D., Vázquez-Abad, M. & Denton, C.P. (2011). Modified-release sildenafil reduces Raynaud's phenomenon attack frequency in limited cutaneous systemic sclerosis. *Arthritis & Rheumatism.* 63 (3): 775-782.

Hewlett, S., Ambler, N., Almeida, C., Cliss, A., Hammond, A., Kitchen, K., Knops, B., Pope, D., Spears, M., Swinkels, A. & Pollock, J. (2011). Self-management of fatigue in rheumatoid arthritis: a randomised controlled trial of group cognitive-behavioural therapy. *Annals of the Rheumatic Diseases.* 70 (6): 1060-1067.

Holsti, L., Oberlander, T.F. & Brant, R. (2011). Does breastfeeding reduce acute procedural pain in preterm infants in the neonatal intensive care unit? A randomized clinical trial. *Pain.* 152 (11): 2575-2581.

Inman, R.D. &Maksymowych, W.P. (2010). A double-blind, placebo-controlled trial of low dose infliximab in ankylosing spondylitis. *The Journal of Rheumatology.* 37 (6): 1203-1210.

Jane, S., Chen, S., Wilkie, D.J., Lin, Y., Foreman, S.W., Beaton, R.D., Fan, J., Lu, M., Wang, Y., Lin, Y. & Liao, M. (2011). Effects of massage on pain, mood status, relaxation, and sleep in Taiwanese patients with metastatic bone pain: A randomized clinical trial. *Pain.* 152 (10): 2432-2442.

Jenks, K., Stebbings, S., Burton, J., Schultz, M., Herbison, P. & Highton, J. (2010). Probiotic therapy for the treatment of spondyloarthritis: a randomized controlled trial. *The Journal of Rheumatology.* 37 (10): 2118-2125.

Jones, G., Sebba, A., Gu, J., Lowenstein, M.B., Calvo, A., Gomez-Reino, J.J., Siri, D.A., Tomšič, M., Alecock, E., Woodworth, T. & Genovese, M.C. (2010). Comparison of tocilizumab monotherapy versus methotrexate monotherapy in patients with moderate to severe rheumatoid arthritis: the AMBITION study. *Annals of the Rheumatic Diseases.* 69 (01): 88-96.

Katz, N., Borenstein, D.G., Birbara, C., Bramson, C., Nemeth, M.A., Smith, M.D. & Brown, M.T. (2011). Efficacy and safety of tanezumab in the treatment of chronic low back pain. *Pain.* 152 (10): 2248-2258.

Keefe, F.J., Shelby, R.A., Somers, T.J., Varia, I., Blazing, M., Waters, S.J., McKee, D., Silva, S., She, L., Blumenthal, J.A., O'Connor, J., Knowles, V., Johnson, P. & Bradley, L. (2011). Effects of coping skills training and sertraline in patients with non-cardiac chest pain: A randomized controlled study. *Pain.* 152 (4): 730-741.

Kemp, S., Roberts, I., Gamble, C., Wilkinson, S., Davidson, J.E., Baildam, E.M., Cleary, A.G., McCann, L.J. & Beresford, M.W. (2010). A randomized comparative trial of generalized vs targeted physiotherapy in the management of childhood hypermobility. *Rheumatology.* 49 (2): 315-325.

Kim, J.S., Bashford, G., Murphy, T.K., Martin, A., Dror, V. & Cheung, R. (2011). Safety and efficacy of pregabalin in patients with central post-stroke pain. *Pain.* 152 (5): 1018-1023.

Kjeken, I., Darre, S., Smedslund, G., Hagen, K.B. & Nossum, R. (2011). Effect of assistive technology in hand osteoarthritis: a randomised controlled trial. *Annals of the Rheumatic Diseases.* 70 (8): 1447-1452.

Kravitz, R.L., Tancredi, D.J., Grennan, T., Kalauokalani, D., Street Jr., R.L., Slee, C.K., Wun, T., Oliver, J.W., Lorig, K. & Franks, P. (2011). Cancer Health Empowerment for Living without Pain (Ca-HELP): effects of a tailored education and coaching intervention on pain and impairment. *Pain.* 152 (7): 1572-1582.

Kremer, J., Ritchlin, C., Mendelsohn, A., Baker, D., Kim, L., Xu, Z., Han, J. & Taylor, P. (2010). Golimumab, a new human anti?tumor necrosis factor ? antibody, administered intravenously in patients with active rheumatoid arthritis: Forty-eight?week efficacy and safety results of a phase III randomized, double-blind, placebo-controlled study. *Arthritis & Rheumatism.* 62 (4): 917-928.

Kume, K., Amano, K., Yamada, S., Hatta, K., Ohta, H. & Kuwaba, N. (2011). Tocilizumab monotherapy reduces arterial stiffness as effectively as etanercept or adalimumab monotherapy in rheumatoid arthritis: an open-label randomized controlled trial. *The Journal of Rheumatology.* 38 (10): 2169-2171.

Litt, M.D., Shafer, D.M. & Kreutzer, D.L. (2010). Brief cognitive-behavioral treatment for TMD pain: Long-term outcomes and moderators of treatment. *Pain.* 151 (1): 110-116.

Lumley, M.A., Leisen, J.C.C., Partridge, R.T., Meyer, T.M., Radcliffe, A.M., Macklem, D.J., Naoum, L.A., Cohen, J.L., Lasichak, L.M., Lubetsky, M.R., Mosley-Williams, A. & Granda, J.L. (2011). Does emotional disclosure about stress improve health in rheumatoid arthritis? Randomized, controlled trials of written and spoken disclosure. *Pain.* 152 (4): 866-877.

Machold, K.P., Landewé, R., Smolen, J.S., Stamm, T.A., van der Heijde, D.M., Verpoort, K.N., Brickmann, K., Vázquez-Mellado, J., Karateev, D.E., Breedveld, F.C., Emery, P. & Huizinga, T.W.J. (2010). The Stop Arthritis Very Early (SAVE) trial, an international multicentre, randomised, double-blind, placebo-controlled trial on glucocorticoids in very early arthritis. *Annals of the Rheumatic Diseases.* 69 (3): 495-502.

Masiero, S., Bonaldo, L., Pigatto, M., Lo Nigro, A., Ramonda, R. & Punzi, L. (2011). Rehabilitation treatment in patients with ankylosing spondylitis stabilized with tumor necrosis factor inhibitor therapy: a randomized controlled trial. *The Journal of Rheumatology.* 38 (7): 1335-1342.

Mease, P.J., Cohen, S., Gaylis, N.B., Chubick, A., Kaell, A.T., Greenwald, M., Agarwal, S., Yin, M. & Kelman, A. (2010). Efficacy and safety of retreatment in patients with rheumatoid arthritis with previous inadequate response to tumor necrosis factor inhibitors: results from the sunrise trial. *The Journal of Rheumatology.* 37 (5): 917-927.

Merrill, J.T., Neuwelt, C.M., Wallace, D.J., Shanahan, J.C., Latinis, K.M., Oates, J.C., Utset, T.O., Gordon, C., Isenberg, D.A., Hsieh, H., Zhang, D. & Brunetta, P.G. (2010). Efficacy and safety of rituximab in moderately-to-severely active systemic lupus erythematosus: The randomized, double-blind, phase ii/iii systemic lupus erythematosus evaluation of rituximab trial. *Arthritis & Rheumatism.* 62 (1): 222-233.

Mok, C.C., Ying, K.Y., To, C.H., Ho, L.Y., Yu, K.L., Lee, H.K. & Ma, K.M. (2011). Raloxifene for prevention of glucocorticoid-induced bone loss: a 12-month randomised double-blinded placebo-controlled trial. *Annals of the Rheumatic Diseases.* 70 (5): 778-784.

Molsberger, A.F., Schneider, T., Gotthardt, H. & Drabik, A. (2010). German Randomized Acupuncture Trial for chronic shoulder pain (GRASP) – A pragmatic, controlled, patient-blinded, multi-centre trial in an outpatient care environment. *Pain.* 151 (1): 146-154.

Munteanu, S.E., Zammit, G.V., Menz, H.B., Landorf, K.B., Handley, C.J., Elzarka, A. & DeLuca, J. (2011). Effectiveness of intra-articular hyaluronan (Synvisc, hylan G-F 20) for the treatment of first metatarsophalangeal joint osteoarthritis: a randomised placebo-controlled trial. *Annals of the Rheumatic Diseases.* 70 (10): 1838-1841.

Navarro-Sarabia, F., Coronel, P., Collantes, E., Navarro, F.J., de la Serna, A.R., Naranjo, A., Gimeno, M. & Herrero-Beaumont, G. (2011). A 40-month multicentre, randomised placebo-controlled study to assess the efficacy and carry-over effect of repeated intra-articular injections of hyaluronic acid in knee osteoarthritis: the AMELIA project. *Annals of the Rheumatic Diseases.* 70 (11): 1957-1962.

Navarro-Sarabia, F., Fernández-Sueiro, J.L., Torre-Alonso, J.C., Gratacos, J., Queiro, R., Gonzalez, C., Loza, E., Linares, L., Zarco, P., Juanola, X., Román-Ivorra, J., Martín-Mola, E., Sanmartí, R., Mulero, J., Diaz, G., Armendáriz, Y. & Collantes, E. (2011). High-dose etanercept in

ankylosing spondylitis: results of a 12-week randomized, double blind, controlled multicentre study (LOADET study). *Rheumatology.* 50 (10): 1828-1837.

Oldenmenger, W.H., Sillevis Smitt, P.A.E., van Montfort, C.A.G.M., de Raaf, P.J. & van der Rijt, C.C.D. (2011). A combined pain consultation and pain education program decreases average and current pain and decreases interference in daily life by pain in oncology outpatients: A randomized controlled trial. *Pain.* 152 (11): 2632-2639.

Pauer, L., Winkelmann, A., Arsenault, P., Jespersen, A., Whelan, L., Atkinson, G., Leon, T. & Zeiher, B. (2011). An international, randomized, double-blind, placebo-controlled, phase iii trial of pregabalin monotherapy in treatment of patients with fibromyalgia. *The Journal of Rheumatology.* 38 (12): 2643-2652.

Peng, B., Pang, X., Wu, Y., Zhao, C. & Song, X. (2010). A randomized placebo-controlled trial of intradiscal methylene blue injection for the treatment of chronic discogenic low back pain. *Pain.* 149 (1): 124-129.

Petri, M.A., Kiani, A.N., Post, W., Christopher-Stine, L. & Magder, L.S. (2011). Lupus Atherosclerosis Prevention Study (LAPS). *Annals of the Rheumatic Diseases.* 70 (5): 760-765.

Petri, M., Brodsky, R.A., Jones, R.J., Gladstone, D., Fillius, M. & Magder, L.S. (2010). High-dose cyclophosphamide versus monthly intravenous cyclophosphamide for systemic lupus erythematosus: A prospective randomized trial. *Arthritis & Rheumatism.* 62 (5): 1487-1493.

Platon, B., Andrell, P., Raner, C., Rudolph, M., Dvoretsky, A. & Mannheimer, C. (2010). High-frequency, high-intensity transcutaneous electrical nerve stimulation as treatment of pain after surgical abortion. *Pain.* 148 (1): 114-119.

Rubbert-Roth, A., Tak, P.P., Zerbini, C., Tremblay, J., Carreño, L., Armstrong, G., Collinson, N. & Shaw, T.M. (2010). Efficacy and safety of various repeat treatment dosing regimens of rituximab in patients with active rheumatoid arthritis: results of a Phase III randomized study (MIRROR). *Rheumatology.* 49 (9): 1683-1693.

Russell, I.J., Holman, A.J., Swick, T.J., Alvarez-Horine, S., Wang, Y.G. & Guinta, D. (2011). Sodium oxybate reduces pain, fatigue, and sleep disturbance and improves functionality in fibromyalgia: results from a 14-week, randomized, double-blind, placebo-controlled study. *Pain.* 152 (5): 1007-1017.

Schmidt, S., Grossman, P., Schwarzer, B., Jena, S., Naumann, J. & Walach, H. (2011). Treating fibromyalgia with mindfulness-based stress reduction: results from a 3-armed randomized controlled trial. *Pain.* 152 (2): 361-369.

Seibold, J.R., Denton, C.P., Furst, D.E., Guillevin, L., Rubin, L.J., Wells, A., Matucci Cerinic, M., Riemekasten, G., Emery, P., Chadha-Boreham, H., Charef, P., Roux, S. & Black, C.M. (2010). Randomized, prospective, placebo-controlled trial of bosentan in interstitial lung disease secondary to systemic sclerosis. *Arthritis & Rheumatism.* 62 (7): 2101-2108.

Sibbitt, W.L., Band, P.A., Chavez-Chiang, N.R., Delea, S.L., Norton, H.E. & Bankhurst, A.D. (2011). A randomized controlled trial of the cost-effectiveness of ultrasound-guided intraarticular injection of inflammatory arthritis. *The Journal of Rheumatology.* 38 (2): 252-263.

Singh, J.A., Mahowald, M.L. & Noorbaloochi, S. (2010). Intraarticular botulinum toxin a for refractory painful total knee arthroplasty: a randomized controlled trial. *The Journal of Rheumatology.* 37 (11): 2377-2386.

Snijders, G.F., van den Ende, C.H., van Riel, P.L., van den Hoogen, F.H. & den Broeder, A.A. (2011). The effects of doxycycline on reducing symptoms in knee osteoarthritis: results from a triple-blinded randomised controlled trial. *Annals of the Rheumatic Diseases.* 70 (7): 1191-1196.

Song, I., Hermann, K., Haibel, H., Althoff, C., Listing, J., Burmester, G., Krause, A., Bohl-Bühler, M., Freundlich, B., Rudwaleit, M. & Sieper, J. (2011). Effects of etanercept versus sulfasalazine in early axial spondyloarthritis on active inflammatory lesions as detected by whole-body MRI (ESTHER): a 48-week randomised controlled trial. *Annals of the Rheumatic Diseases.* 70 (4): 590-596.

Strand, V., Simon, L.S., Dougados, M., Sands, G.H., Bhadra, P., Breazna, A. & Immitt, J. (2011). Treatment of osteoarthritis with continuous versus intermittent Celecoxib. *The Journal of Rheumatology.* 38 (12): 2625-2634.

Tak, P.P., Rigby, W.F., Rubbert-Roth, A., Peterfy, C.G., van Vollenhoven, R.F., Stohl, W., Hessey, E., Chen, A., Tyrrell, H., Shaw, T.M. & for the IMAGE Investigators (2011). Inhibition of joint damage and improved clinical outcomes with rituximab plus methotrexate in early active rheumatoid arthritis: the IMAGE trial. *Annals of the Rheumatic Diseases.* 70 (1): 39-46.

Tanaka, Y., Harigai, M., Takeuchi, T., Yamanaka, H., Ishiguro, N., Yamamoto, K., Miyasaka, N., Koike, T., Kanazawa, M., Oba, T., Yoshinari, T. & Baker, D. (2011). Golimumab in combination with methotrexate in Japanese patients with active rheumatoid arthritis: results of the GO-FORTH study. *Annals of the Rheumatic Diseases.*

Taylor, P.C., Quattrocchi, E., Mallett, S., Kurrasch, R., Petersen, J. & Chang, D.J. (2011). Ofatumumab, a fully human anti-CD20 monoclonal antibody, in biological-naive, rheumatoid arthritis patients with an inadequate response to methotrexate: a randomised, double-blind, placebo-controlled clinical trial. *Annals of the Rheumatic Diseases.* 70 (12): 2119-2125.

Thorn, B.E., Day, M.A., Burns, J., Kuhajda, M.C., Gaskins, S.W., Sweeney, K., McConley, R., Ward, L.C. & Cabbil, C. (2011). Randomized trial of group cognitive behavioral therapy compared with a pain education control for low-literacy rural people with chronic pain. *Pain.* 152 (12): 2710-2720.

Turner, J.A., Mancl, L., Huggins, K.H., Sherman, J.J., Lentz, G. & LeResche, L. (2011). Targeting temporomandibular disorder pain treatment to hormonal fluctuations: A randomized clinical trial. *Pain.* 152 (9): 2074-2084.

Tynjälä, P., Vähäsalo, P., Tarkiainen, M., Kröger, L., Aalto, K., Malin, M., Putto-Laurila, A., Honkanen, V. & Lahdenne, P. (2011). Aggressive Combination Drug Therapy in Very Early Polyarticular Juvenile Idiopathic Arthritis (ACUTE–JIA): a multicentre randomised open-label clinical trial. *Annals of the Rheumatic Diseases.* 70 (9): 1605-1612.

Urata, Y., Uesato, R., Tanaka, D., Nakamura, Y. & Motomura, S. (2011). Treating to target matrix metalloproteinase 3 normalisation together with disease activity score below 2.6 yields better effects than each alone in rheumatoid arthritis patients: T-4 Study. *Annals of the Rheumatic Diseases.*

Verbruggen, G., Wittoek, R., Cruyssen, B.V. & Elewaut, D. (2011). Tumour necrosis factor blockade for the treatment of erosive osteoarthritis of the interphalangeal finger joints: a double blind, randomised trial on structure modification. *Annals of the Rheumatic Diseases.*

Verstappen, S.M.M., McCoy, M.J., Roberts, C., Dale, N.E., Hassell, A.B. & Symmons, D.P.M. (2010). Beneficial effects of a 3-week course of intramuscular glucocorticoid injections in patients with very early inflammatory polyarthritis: results of the STIVEA trial. *Annals of the Rheumatic Diseases.* 69 (3): 503-509.

Wang, L., Zhang, X., Guo, J., Liu, H., Zhang, Y., Liu, C., Yi, J., Wang, L., Zhao, J. & Li, S. (2011). Efficacy of acupuncture for migraine prophylaxis: A single-blinded, double-dummy, randomized controlled trial. *Pain.* 152 (8): 1864-1871.

Wetherell, J.L., Afari, N., Rutledge, T., Sorrell, J.T., Stoddard, J.A., Petkus, A.J., Solomon, B.C., Lehman, D.H., Liu, L., Lang, A.J. & Hampton Atkinson, J. (2011). A randomized, controlled trial of acceptance and commitment therapy and cognitive-behavioral therapy for chronic pain. *Pain.* 152 (9): 2098-2107.

Williams, D.A., Kuper, D., Segar, M., Mohan, N., Sheth, M. & Clauw, D.J. (2010). Internet-enhanced management of fibromyalgia: A randomized controlled trial. *Pain.* 151 (3): 694-702.

Zangi, H.A., Mowinckel, P., Finset, A., Eriksson, L.R., Høystad, T.Ø., Lunde, A.K. & Hagen, K.B. (2011). A mindfulness-based group intervention to reduce psychological distress and fatigue in patients with inflammatory rheumatic joint diseases: a randomised controlled trial. *Annals of the Rheumatic Diseases.*

Zayat, A.S., Conaghan, P.G., Sharif, M., Freeston, J.E., Wenham, C., Hensor, E.M.A., Emery, P. & Wakefield, R.J. (2011). Do non-steroidal anti-inflammatory drugs have a significant effect on detection and grading of ultrasound-detected synovitis in patients with rheumatoid arthritis? Results from a randomised study. *Annals of the Rheumatic Diseases.* 70 (10): 1746-1751.

Appendix 2: Systematic review - data extraction form

A. Basic details

1. Serial Number:
2. Author:
3. Year of publication:
4. Journal Name:
5. Disease category being studied:
6. Number of groups compared:
7. Primary objective:
8. Primary outcome – variable type [1-numerical (discrete/continuous); 2-categorical (nominal/ordinal)]
    a. As measured :
    b. As analysed :
9. Design of trial (1- cluster RCT; 2 – other):
10. Multicentre trial? (1 – yes; 0 - no):
11. Number of follow-up visit (1 – 1 follow-up; 2 – 2 or more follow-up):
12. Sample size
    a. Calculated sample size:
    b. Adjustment for attrition in sample size calculation (1 – yes; 0 – no):
    c. Number of subjects randomized:

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total

13. Number of participants completed the trial (at the primary endpoint)

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total

14. Reported analysis strategy:

B. Analysis strategy (primary analysis)

1. Size of the trial (Number of participants in the analysis):

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total

2. Handling of ineligible subjects (Subsequent ineligible subjects - based on inclusion or exclusion criteria)

    a. Presence of ineligible subjects (1 – yes; 0 – no):

    b. If yes in (a), how the subjects were handled (1 -excluded; 2 – withdrawal; 3 – included; 99 –not clear): *[withdrawal (treated as missing data) can be either patient's or clinician's decision]*

    c. If excluded, how many subjects

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total

3. Handling of major treatment protocol violations

3.1. Presence of treatment crossover

    a. Treatment crossover (1 – yes; 0 – no):

    b. If yes, analysed as (1 – excluded; 2 – randomized; 3 - treated; 99- not clear):

    c. If excluded, how many subjects

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total

3.2. Major treatment protocol violation (other than treatment crossover & ineligible inclusion) [for e.g., subjects may not have followed the treatment procedure properly or may have taken other medication along with assigned treatment, etc.]

    a. Whether major protocol violations were reported? (1 – yes; 0 – no):

    b. If yes in (a), how the subjects were handled (1 – excluded; 2 – withdrawal; 3 – included; 99 –not clear):

    c. If excluded, how many subjects

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total

4. Handling of missing data

4.1. Subjects were randomized, but not started the treatment [or discontinued before start of the treatment]

a. Subjects with that condition (1 –present; 2 – Absent):

b. If present, how many subjects

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total

c. If present, how the subjects were handled (1 – excluded; 2 – included; 99 –not clear):

d. If included, method of analysing missing data:

4.2. Treatment was started, but no post-baseline measurements were available [Treatment was started, but discontinued prior to the post-baseline measurement]

a. Subjects with that condition (1 –present; 2 – Absent):

b. If present, how many subjects

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total

c. If present, how the subjects were handled (1 – excluded; 2 – included; 99 –not clear):

d. If included, method of handling missing data:

4.3. Missing at primary endpoint (excluding subjects with the above conditions 4.1 & 4.2) [Post-baseline measurements were available, but discontinued prior to the primary endpoint]

a. Missing data (1 –present; 2 – Absent):

b. If present, how many subjects

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total

c. If present, how the subjects were handled (1 – excluded; 2 – included; 99 –not clear):

d. If included, method of handling missing data:

5. Reason for missing data (in situations 4.2 & 4.3)

a. Major reason for patient withdrawal:

(1 – Adverse event; 2 – inefficacy; 3- lost to follow-up; 4- other; 99- not clear)

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total

6. Analysis method (Hypothesis testing)

a. Analysis method:

7. Sensitivity analysis (with alternative assumptions made about missing data)

    a. Whether sensitivity analysis was performed (1 – yes; 0 – no):

    b. If yes in (a), analysis result was presented (1 –yes; 0 – no):

    c. If yes in (a), handling of subjects with missing data (1 – excluded; 2 – included; 99 –not clear):

    d. If included, method of handling missing data:

    e. Analysis method – hypothesis testing:

Appendix 3: Stata syntax for data simulation and analyses

```
/*Smulation program (sim_dropout) to generate based on 30% MCAR missing data under different scenarios in ANCOVA, MMRM(reml), LOCF, & MI. */

clear
capture program drop sim_dropout_30
program define sim_dropout_30, rclass
version 12.0
syntax , TRAJectory(integer) cs(integer) obs(integer)

//Correlation structure (positive semi-definite): strong - 1: weak - 0
if `cs'==0 {
matrix input Corr = (1, 0.45, 0.39, 0.30\0.45, 1.0, 0.41, 0.34\0.39, 0.41, 1.0, 0.40\0.30, 0.34, 0.40, 1.0)
}
if `cs'==1 {
matrix input Corr = (1.0, 0.75, 0.63, 0.54\0.75, 1.0,0.71, 0.59\0.63, 0.71, 1.0, 0.66\0.54, 0.59, 0.66, 1.0)
}
//std deviation: low(=1), medium (=2), high(=3)
if `sd'==1 {
matrix input SD = (10.2, 10.7, 11.4, 12.2)
}
if `sd'==2 {
matrix input SD = (14.1, 14.6, 16.9, 17.7)
}
if `sd'==3 {
matrix input SD = (24.5, 25, 25.7, 26.5)
}
local no=`obs'/2
set obs `no'
 //under trajectory 1:
 if `trajectory'==1 {
matrix input Mean1 = (53, 50, 46, 40) /*Mean for group 0*/
matrix input Mean2 = (53, 50, 46, 40) /*Mean for group 1*/
drawnorm y01 y11 y21 y31, means(Mean1) sds(SD) corr(Corr)
drawnorm y02 y12 y22 y32, means(Mean2) sds(SD) corr(Corr)
 }
 //under trajectory 2:
 if `trajectory'==2 {
matrix input Mean1 = (53, 50, 53, 48) /*Mean for group 0*/
matrix input Mean2 = (53, 31, 28, 30) /*Mean for group 1*/
drawnorm y01 y11 y21 y31, means(Mean1) sds(SD) corr(Corr)
drawnorm y02 y12 y22 y32, means(Mean2) sds(SD) corr(Corr)
 }
 //under trajectory 3: main one
 if `trajectory'==3 {
matrix input Mean1 = (53, 50, 46, 40) /*Mean for group 0*/
matrix input Mean2 = (53, 47, 41, 31) /*Mean for group 1*/
drawnorm y01 y11 y21 y31, means(Mean1) sds(SD) corr(Corr)
drawnorm y02 y12 y22 y32, means(Mean2) sds(SD) corr(Corr)
 }
 //under trajectory 4:
 if `trajectory'==4 {
matrix input Mean1 = (53, 45, 48, 32) /*Mean for group 0*/
matrix input Mean2 = (53, 51, 43, 23) /*Mean for group 1*/
drawnorm y01 y11 y21 y31, means(Mean1) sds(SD) corr(Corr)
```

```
drawnorm y02 y12 y22 y32, means(Mean2) sds(SD) corr(Corr)
 }

stack y01 y11 y21 y31 y02 y12 y22 y32, into(y0 y1 y2 y3) clear
gen group=(_stack==2)
drop _stack
gen id=_n
gen baseline=y0

/*****************************************************************************************************
Generating missing datasets based on 21 dropout pattern (the final dataset will be a combination of
22 datasets);  dr - dropout rate,  dm - dropout mechanism,  dp - direction of dropout
*****************************************************************************************************/
//generating baseline variable for each datasets
gen baseline=y0
foreach dr in 30 {
foreach dm in 01 02 04 06 08 10 12 {
foreach dp in 01 02 03 {
 gen y`dm'`dr'`dp'0=y0
 }
 }
 }

/*****************************************************************************************************
Dropout pattern 01: (MCAR 01; Dropout rate 30%; Dropout direction 01 - equal dropout between
groups)
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/10%; visit 2 - 20%/20%; visit 3 - 30%/30%
*****************************************************************************************************/
//generating new set of variables for dropout pattern 1
gen y0130011=y1
gen y0130012=y2
gen y0130013=y3

**group 0**
//generate 10% dropouts at visit 1 in control group
gen rand=runiform() if group==0 /*create a new variable to randomly order the outcome values at
visit 1*/
_pctile rand if group==0, nq(100)
scalar cutoff=r(r90)
replace y0130011=. if rand >=cutoff & rand<.& group==0 /*10% obns were randomly deleted at
visit 1*/
replace y0130012=. if y0130011==. & group==0 /* corresponding obns were deleted at visit 2*/
replace y0130013=. if y0130011==. & group==0 /* corresponding obns were deleted at visit 3*/
drop rand /*dropped the new variable*/

//generate additional 10% dropouts at visit 2 in control group
gen rand=runiform() if y0130012!=.& group==0 /*create a new variable to randomly order the
remaining outcome values at visit 2*/
_pctile rand if y0130012!=.& group==0 , nq(90)
scalar cutoff=r(r80)
replace y0130012=. if rand >=cutoff & rand<.& group==0 /*10% observations were randomly
deleted at visit 2*/
replace y0130013=. if y0130012==.& group==0 /* corresponding obns were deleted at visit 3*/
drop rand /*dropped the new variable*/

//generate additional 10% dropouts at visit 3 in control group
gen rand=runiform() if y0130013!=.& group==0 /*create a new variable to randomly order the
remaining outcome values at visit 3*/
_pctile rand if y0130013!=.& group==0 , nq(80)
```

```
scalar cutoff=r(r70)
replace y0130013=. if rand >=cutoff & rand<.& group==0 /*10% observations were randomly
deleted at visit 3*/
drop rand /*dropped the new variable*/

**group 1**
 //generate 10% dropouts at visit 1 in experimental group
gen rand=runiform() if group==1
_pctile rand if group==1, nq(100)
scalar cutoff=r(r90)
replace y0130011=. if rand >=cutoff & rand<.& group==1
replace y0130012=. if y0130011==. & group==1
replace y0130013=. if y0130011==. & group==1
drop rand
//generate additional 10% dropouts at visit 2 in experimental group
gen rand=runiform() if y0130012!=.& group==1
_pctile rand if y0130012!=.& group==1 , nq(90)
scalar cutoff=r(r80)
replace y0130012=. if rand >=cutoff & rand<.& group==1
replace y0130013=. if y0130012==.& group==1
drop rand
//generate additional 10% dropouts at visit 3 in experimental group
gen rand=runiform() if y0130013!=.& group==1
_pctile rand if y0130013!=.& group==1 , nq(80)
scalar cutoff=r(r70)
replace y0130013=. if rand >=cutoff & rand<.& group==1
drop rand

/*****************************************************************************************************
Dropout pattern 02: (MCAR 01; Dropout rate 30%; Dropout direction 02 - higher dropout in
experimental group)
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 5%/10%; visit 2 - 15%/25%; visit 3 - 20%/40%
*****************************************************************************************************/
//generating new set of variables for dropout pattern 2
gen y0130021=y1
gen y0130022=y2
gen y0130023=y3
**group 0**
gen rand=runiform() if group==0
_pctile rand if group==0, nq(100)
scalar cutoff=r(r95)
replace y0130021=. if rand >=cutoff & rand<.& group==0
replace y0130022=. if y0130021==. & group==0
replace y0130023=. if y0130021==. & group==0
drop rand
gen rand=runiform() if y0130022!=.& group==0
_pctile rand if y0130022!=.& group==0 , nq(95)
scalar cutoff=r(r85)
replace y0130022=. if rand >=cutoff & rand<.& group==0
replace y0130023=. if y0130022==.& group==0
drop rand
gen rand=runiform() if y0130023!=.& group==0
_pctile rand if y0130023!=.& group==0 , nq(85)
scalar cutoff=r(r80)
replace y0130023=. if rand >=cutoff & rand<.& group==0
drop rand
**group 1**
gen rand=runiform() if group==1
_pctile rand if group==1, nq(100)
```

```
scalar cutoff=r(r90)
replace y0130021=. if rand >=cutoff & rand<.& group==1
replace y0130022=. if y0130021==. & group==1
replace y0130023=. if y0130021==. & group==1
drop rand
gen rand=runiform() if y0130022!=.& group==1
_pctile rand if y0130022!=.& group==1 , nq(90)
scalar cutoff=r(r75)
replace y0130022=. if rand >=cutoff & rand<.& group==1
replace y0130023=. if y0130022==.& group==1
drop rand
gen rand=runiform() if y0130023!=.& group==1
_pctile rand if y0130023!=.& group==1 , nq(75)
scalar cutoff=r(r60)
replace y0130023=. if rand >=cutoff & rand<.& group==1
drop rand

/*****************************************************************************************************
Dropout pattern 03: (MCAR 01; Dropout rate 30%; Dropout direction 03 - high dropout in control group)
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/5%; visit 2 - 25%/15%; visit 3 - 40%/20%
*****************************************************************************************************/
//generating new set of variables for dropout pattern 3
gen y0130031=y1
gen y0130032=y2
gen y0130033=y3
**group 0**
gen rand=runiform() if group==0
_pctile rand if group==0, nq(100)
scalar cutoff=r(r90)
replace y0130031=. if rand >=cutoff & rand<.& group==0
replace y0130032=. if y0130031==. & group==0
replace y0130033=. if y0130031==. & group==0
drop rand
gen rand=runiform() if y0130032!=.& group==0
_pctile rand if y0130032!=.& group==0 , nq(90)
scalar cutoff=r(r75)
replace y0130032=. if rand >=cutoff & rand<.& group==0
replace y0130033=. if y0130032==.& group==0
drop rand
gen rand=runiform() if y0130033!=.& group==0
_pctile rand if y0130033!=.& group==0 , nq(75)
scalar cutoff=r(r60)
replace y0130033=. if rand >=cutoff & rand<.& group==0
drop rand
**group 1**
gen rand=runiform() if group==1
_pctile rand if group==1, nq(100)
scalar cutoff=r(r95)
replace y0130031=. if rand >=cutoff & rand<.& group==1
replace y0130032=. if y0130031==. & group==1
replace y0130033=. if y0130031==. & group==1
drop rand
gen rand=runiform() if y0130032!=.& group==1
_pctile rand if y0130032!=.& group==1 , nq(95)
scalar cutoff=r(r85)
replace y0130032=. if rand >=cutoff & rand<.& group==1
replace y0130033=. if y0130032==.& group==1
drop rand
```

```
gen rand=runiform() if y0130033!=.& group==1
_pctile rand if y0130033!=.& group==1 , nq(85)
scalar cutoff=r(r80)
replace y0130033=. if rand >=cutoff & rand<.& group==1
drop rand

/*****************************************************************************************************
Dropout pattern 04: (MAR 2 - dropout depends on baseline values; Dropout rate 30%; Dropout direction 01 - equal dropout between groups)*
Deletion restricted to subjects with high baseline score in both groups ( i.e., above p50 of y0)
cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/10%; visit 2 - 20%/20%; visit 3 - 30%/30%
*****************************************************************************************************/
//generating new set of variables for dropout pattern 4
gen y0230011=y1
gen y0230012=y2
gen y0230013=y3
**group 0**
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0 /*randomly order the baseline values that are above 50th percentile*/
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y0230011=. if rand >=cutoff & rand<. & group==0 /*randomly delete 10% values among the outcome at visit 1 where the corresponding baseline values are above 50th percentile of y0*/
replace y0230012=. if y0230011==.
replace y0230013=. if y0230011==.
drop rand
gen rand=runiform() if y0230012!=. & y0>=p50 & group==0 /*randomly order the baseline values that are above 50th percentile if the values at subsequent visits are not missing */
_pctile rand if y0230012!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r30)
replace y0230012=. if rand >=cutoff & rand<.& group==0
replace y0230013=. if y0230012==.
drop rand
gen rand=runiform() if y0230013!=. & y0>=p50 & group==0
_pctile rand if y0230013!=. & rand<. & group==0, nq(30)
scalar cutoff=r(r20)
replace y0230013=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y0230011=. if rand >=cutoff & rand<. & group==1
replace y0230012=. if y0230011==.
replace y0230013=. if y0230011==.
drop rand
gen rand=runiform() if y0230012!=. & y0>=p50 & group==1
_pctile rand if y0230012!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r30)
replace y0230012=. if rand >=cutoff & rand<.& group==1
replace y0230013=. if y0230012==.
drop rand
gen rand=runiform() if y0230013!=. & y0>=p50 & group==1
_pctile rand if y0230013!=. & rand<. & group==1, nq(30)
```

```
scalar cutoff=r(r20)
replace y0230013=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
Dropout pattern 05: (MAR 2 - dropout depends on baseline values; Dropout rate 30%; Dropout
direction 02 - higher dropout in group 1)
Deletion restricted to subjects with high baseline score in both groups ( i.e., above p50 of y0)
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 5%/10%; visit 2 - 15%/25%; visit 3 - 20%/40%
*****************************************************************************************************/
//generating new set of variables for dropout pattern 5
gen y0230021=y1
gen y0230022=y2
gen y0230023=y3
**group 0**
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r45)
replace y0230021=. if rand >=cutoff & rand<. & group==0
replace y0230022=. if y0230021==.
replace y0230023=. if y0230021==.
drop rand
gen rand=runiform() if y0230022!=. & y0>=p50 & group==0
_pctile rand if y0230022!=. & rand<. & group==0 , nq(45)
scalar cutoff=r(r35)
replace y0230022=. if rand >=cutoff & rand<.& group==0
replace y0230023=. if y0230022==.
drop rand
gen rand=runiform() if y0230023!=. & y0>=p50 & group==0
_pctile rand if y0230023!=. & rand<. & group==0, nq(35)
scalar cutoff=r(r30)
replace y0230023=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y0230021=. if rand >=cutoff & rand<. & group==1
replace y0230022=. if y0230021==.
replace y0230023=. if y0230021==.
drop rand
gen rand=runiform() if y0230022!=. & y0>=p50 & group==1
_pctile rand if y0230022!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r25)
replace y0230022=. if rand >=cutoff & rand<.& group==1
replace y0230023=. if y0230022==.
drop rand
gen rand=runiform() if y0230023!=. & y0>=p50 & group==1
_pctile rand if y0230023!=. & rand<. & group==1, nq(25)
scalar cutoff=r(r10)
replace y0230023=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
```

```
Dropout pattern 06: (MAR 2 - dropout depends on baseline values; Dropout rate 30%; Dropout
direction 03 - higher dropout in group 0)
Deletion restricted to subjects with high baseline score in both groups ( i.e., above p50 of y0)
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/5%; visit 2 - 25%/15%; visit 3 - 40%/20%
*****************************************************************************************************/
//generating new set of variables for dropout pattern 06
gen y0230031=y1
gen y0230032=y2
gen y0230033=y3
**group 0**
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y0230031=. if rand >=cutoff & rand<. & group==0
replace y0230032=. if y0230031==.
replace y0230033=. if y0230031==.
drop rand
gen rand=runiform() if y0230032!=. & y0>=p50 & group==0
_pctile rand if y0230032!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r25)
replace y0230032=. if rand >=cutoff & rand<.& group==0
replace y0230033=. if y0230032==.
drop rand
gen rand=runiform() if y0230033!=. & y0>=p50 & group==0
_pctile rand if y0230033!=. & rand<. & group==0, nq(25)
scalar cutoff=r(r10)
replace y0230033=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r45)
replace y0230031=. if rand >=cutoff & rand<. & group==1
replace y0230032=. if y0230031==.
replace y0230033=. if y0230031==.
drop rand
gen rand=runiform() if y0230032!=. & y0>=p50 & group==1
_pctile rand if y0230032!=. & rand<. & group==1 , nq(45)
scalar cutoff=r(r35)
replace y0230032=. if rand >=cutoff & rand<.& group==1
replace y0230033=. if y0230032==.
drop rand
gen rand=runiform() if y0230033!=. & y0>=p50 & group==1
_pctile rand if y0230033!=. & rand<. & group==1, nq(35)
scalar cutoff=r(r30)
replace y0230033=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
Dropout pattern 07: (MAR 4 - dropout depends on baseline values; Dropout rate 30%; Dropout
direction 01 - equal dropout between groups)
Deletion restricted to subjects with high baseline score in control group and subjects with low
baseline score in experimental group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/10%; visit 2 - 20%/20%; visit 3 - 30%/30%
```

```
**********************************************************************************************************/
gen y0430011=y1
gen y0430012=y2
gen y0430013=y3
**group 0**
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y0430011=. if rand >=cutoff & rand<. & group==0
replace y0430012=. if y0430011==.
replace y0430013=. if y0430011==.
drop rand
gen rand=runiform() if y0430012!=. & y0>=p50 & group==0
_pctile rand if y0430012!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r30)
replace y0430012=. if rand >=cutoff & rand<.& group==0
replace y0430013=. if y0430012==.
drop rand
gen rand=runiform() if y0430013!=. & y0>=p50 & group==0
_pctile rand if y0430013!=. & rand<. & group==0, nq(30)
scalar cutoff=r(r20)
replace y0430013=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0<=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y0430011=. if rand >=cutoff & rand<. & group==1
replace y0430012=. if y0430011==.
replace y0430013=. if y0430011==.
drop rand
gen rand=runiform() if y0430012!=. & y0<=p50 & group==1
_pctile rand if y0430012!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r30)
replace y0430012=. if rand >=cutoff & rand<.& group==1
replace y0430013=. if y0430012==.
drop rand
gen rand=runiform() if y0430013!=. & y0<=p50 & group==1
_pctile rand if y0430013!=. & rand<. & group==1, nq(30)
scalar cutoff=r(r20)
replace y0430013=. if rand >=cutoff & rand<. & group==1
drop rand

/**********************************************************************************************************
Dropout pattern 08: (MAR 4 - dropout depends on baseline values; Dropout rate 30%; Dropout direction 02 - higher dropout rate in group 1)
Deletion restricted to subjects with high baseline score in control group and subjects with low baseline score in experimental group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 5%/10%; visit 2 - 15%/25%; visit 3 - 20%/40%
**********************************************************************************************************/
gen y0430021=y1
gen y0430022=y2
gen y0430023=y3
**group 0**
```

```
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r45)
replace y0430021=. if rand >=cutoff & rand<. & group==0
replace y0430022=. if y0430021==.
replace y0430023=. if y0430021==.
drop rand
gen rand=runiform() if y0430022!=. & y0>=p50 & group==0
_pctile rand if y0430022!=. & rand<. & group==0 , nq(45)
scalar cutoff=r(r35)
replace y0430022=. if rand >=cutoff & rand<.& group==0
replace y0430023=. if y0430022==.
drop rand
gen rand=runiform() if y0430023!=. & y0>=p50 & group==0
_pctile rand if y0430023!=. & rand<. & group==0, nq(35)
scalar cutoff=r(r30)
replace y0430023=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0<=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y0430021=. if rand >=cutoff & rand<. & group==1
replace y0430022=. if y0430021==.
replace y0430023=. if y0430021==.
drop rand
gen rand=runiform() if y0430022!=. & y0<=p50 & group==1
_pctile rand if y0430022!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r25)
replace y0430022=. if rand >=cutoff & rand<.& group==1
replace y0430023=. if y0430022==.
drop rand
gen rand=runiform() if y0430023!=. & y0<=p50 & group==1
_pctile rand if y0430023!=. & rand<. & group==1, nq(25)
scalar cutoff=r(r10)
replace y0430023=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
Dropout pattern 09: (MAR 4 - dropout depends on baseline values; Dropout rate 30%; Dropout
direction 03 - higher dropout rate in group 0)
Deletion restricted to subjects with high baseline score in control group and subjects with low
baseline score in experimental group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/5%; visit 2 - 25%/15%; visit 3 - 40%/20%
*****************************************************************************************************/
gen y0430031=y1
gen y0430032=y2
gen y0430033=y3
**group 0**
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
```

```
replace y0430031=. if rand >=cutoff & rand<. & group==0
replace y0430032=. if y0430031==.
replace y0430033=. if y0430031==.
drop rand
gen rand=runiform() if y0430032!=. & y0>=p50 & group==0
_pctile rand if y0430032!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r25)
replace y0430032=. if rand >=cutoff & rand<.& group==0
replace y0430033=. if y0430032==.
drop rand
gen rand=runiform() if y0430033!=. & y0>=p50 & group==0
_pctile rand if y0430033!=. & rand<. & group==0, nq(25)
scalar cutoff=r(r10)
replace y0430033=. if rand >=cutoff & rand<. & group==0
drop rand
**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0<=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r45)
replace y0430031=. if rand >=cutoff & rand<. & group==1
replace y0430032=. if y0430031==.
replace y0430033=. if y0430031==.
drop rand
gen rand=runiform() if y0430032!=. & y0<=p50 & group==1
_pctile rand if y0430032!=. & rand<. & group==1 , nq(45)
scalar cutoff=r(r35)
replace y0430032=. if rand >=cutoff & rand<.& group==1
replace y0430033=. if y0430032==.
drop rand
gen rand=runiform() if y0430033!=. & y0<=p50 & group==1
_pctile rand if y0430033!=. & rand<. & group==1, nq(35)
scalar cutoff=r(r30)
replace y0430033=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
Dropout pattern 10: (MAR 06 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 01 - equal dropout between groups)
Deletion restricted to subjects with high last follow-up score in both groups ( i.e., above p50 of last
follow-up visit)
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/10%; visit 2 - 20%/20%; visit 3 - 30%/30%
*****************************************************************************************************/
gen y0630011=y1
gen y0630012=y2
gen y0630013=y3
**group 0**
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y0630011=. if rand >=cutoff & rand<. & group==0
replace y0630012=. if y0630011==.
replace y0630013=. if y0630011==.
drop rand

_pctile y0630011 if group==0, nq(90)
```

```
scalar p50=r(r50)
gen rand=runiform() if y0630012!=. & y0630011>=p50 & group==0
_pctile rand if y0630012!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r30)
replace y0630012=. if rand >=cutoff & rand<.& group==0
replace y0630013=. if y0630012==.
drop rand

_pctile y0630012 if group==0, nq(80)
scalar p50=r(r50)
gen rand=runiform() if y0630013!=. & y0630012>=p50 & group==0
_pctile rand if y0630013!=. & rand<. & group==0, nq(30)
scalar cutoff=r(r20)
replace y0630013=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y0630011=. if rand >=cutoff & rand<. & group==1
replace y0630012=. if y0630011==.
replace y0630013=. if y0630011==.
drop rand

_pctile y0630011 if group==1, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y0630012!=. & y0630011>=p50 & group==1
_pctile rand if y0630012!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r30)
replace y0630012=. if rand >=cutoff & rand<.& group==1
replace y0630013=. if y0630012==.
drop rand

_pctile y0630012 if group==1, nq(80)
scalar p50=r(r50)
gen rand=runiform() if y0630013!=. & y0630012>=p50 & group==1
_pctile rand if y0630013!=. & rand<. & group==1, nq(30)
scalar cutoff=r(r20)
replace y0630013=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************************
Dropout pattern 11: (MAR 06 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 02 - higher dropout in group 1)
Deletion restricted to subjects with high last follow-up score in both groups ( i.e., above p50 of last
follow-up visit)
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 5%/10%; visit 2 - 15%/25%; visit 3 - 20%/40%
*****************************************************************************************************************/
gen y0630021=y1
gen y0630022=y2
gen y0630023=y3
**group 0**
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
```

```
scalar cutoff=r(r45)
replace y0630021=. if rand >=cutoff & rand<. & group==0
replace y0630022=. if y0630021==.
replace y0630023=. if y0630021==.
drop rand
_pctile y0630021 if group==0, nq(95)
scalar p50=r(r50)
gen rand=runiform() if y0630022!=. & y0630021>=p50 & group==0
_pctile rand if y0630022!=. & rand<. & group==0 , nq(45)
scalar cutoff=r(r35)
replace y0630022=. if rand >=cutoff & rand<.& group==0
replace y0630023=. if y0630022==.
drop rand
_pctile y0630022 if group==0, nq(85)
scalar p50=r(r50)
gen rand=runiform() if y0630023!=. & y0630022>=p50 & group==0
_pctile rand if y0630023!=. & rand<. & group==0, nq(35)
scalar cutoff=r(r30)
replace y0630023=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y0630021=. if rand >=cutoff & rand<. & group==1
replace y0630022=. if y0630021==.
replace y0630023=. if y0630021==.
drop rand
_pctile y0630021 if group==1, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y0630022!=. & y0630021>=p50 & group==1
_pctile rand if y0630022!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r25)
replace y0630022=. if rand >=cutoff & rand<.& group==1
replace y0630023=. if y0630022==.
drop rand
_pctile y0630022 if group==1, nq(75)
scalar p50=r(r50)
gen rand=runiform() if y0630023!=. & y0630022>=p50 & group==1
_pctile rand if y0630023!=. & rand<. & group==1, nq(25)
scalar cutoff=r(r10)
replace y0630023=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
Dropout pattern 12: (MAR 06 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 03 - higher dropout in group 0)
Deletion restricted to subjects with high last follow-up score in both groups ( i.e., above p50 of last
follow-up visit)
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/5%; visit 2 - 25%/15%; visit 3 - 40%/20%
*****************************************************************************************************/
gen y0630031=y1
gen y0630032=y2
gen y0630033=y3

**group 0**
```

```
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y0630031=. if rand >=cutoff & rand<. & group==0
replace y0630032=. if y0630031==.
replace y0630033=. if y0630031==.
drop rand
_pctile y0630031 if group==0, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y0630032!=. & y0630031>=p50 & group==0
_pctile rand if y0630032!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r25)
replace y0630032=. if rand >=cutoff & rand<.& group==0
replace y0630033=. if y0630032==.
drop rand
_pctile y0630032 if group==0, nq(75)
scalar p50=r(r50)
gen rand=runiform() if y0630033!=. & y0630032>=p50 & group==0
_pctile rand if y0630033!=. & rand<. & group==0, nq(25)
scalar cutoff=r(r10)
replace y0630033=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r45)
replace y0630031=. if rand >=cutoff & rand<. & group==1
replace y0630032=. if y0630031==.
replace y0630033=. if y0630031==.
drop rand
_pctile y0630031 if group==1, nq(95)
scalar p50=r(r50)
gen rand=runiform() if y0630032!=. & y0630031>=p50 & group==1
_pctile rand if y0630032!=. & rand<. & group==1 , nq(45)
scalar cutoff=r(r35)
replace y0630032=. if rand >=cutoff & rand<.& group==1
replace y0630033=. if y0630032==.
drop rand
_pctile y0630032 if group==1, nq(85)
scalar p50=r(r50)
gen rand=runiform() if y0630033!=. & y0630032>=p50 & group==1
_pctile rand if y0630033!=. & rand<. & group==1, nq(35)
scalar cutoff=r(r30)
replace y0630033=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
Dropout pattern 13: (MAR 08 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 01 - equal dropout between groups)
Deletion restricted to subjects with high last follow-up score in control group and low score in
experimental group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/10%; visit 2 - 20%/20%; visit 3 - 30%/30%*/
*****************************************************************************************************/
gen y0830011=y1
```

```
gen y0830012=y2
gen y0830013=y3

**group 0**
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y0830011=. if rand >=cutoff & rand<. & group==0
replace y0830012=. if y0830011==.
replace y0830013=. if y0830011==.
drop rand
_pctile y0830011 if group==0, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y0830012!=. & y0830011>=p50 & group==0
_pctile rand if y0830012!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r30)
replace y0830012=. if rand >=cutoff & rand<.& group==0
replace y0830013=. if y0830012==.
drop rand
_pctile y0830012 if group==0, nq(80)
scalar p50=r(r50)
gen rand=runiform() if y0830013!=. & y0830012>=p50 & group==0
_pctile rand if y0830013!=. & rand<. & group==0, nq(30)
scalar cutoff=r(r20)
replace y0830013=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0<=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y0830011=. if rand >=cutoff & rand<. & group==1
replace y0830012=. if y0830011==.
replace y0830013=. if y0830011==.
drop rand
_pctile y0830011 if group==1, nq(90)
scalar p50=r(r40)
gen rand=runiform() if y0830012!=. & y0830011<=p50 & group==1
_pctile rand if y0830012!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r30)
replace y0830012=. if rand >=cutoff & rand<.& group==1
replace y0830013=. if y0830012==.
drop rand
_pctile y0830012 if group==1, nq(80)
scalar p50=r(r30)
gen rand=runiform() if y0830013!=. & y0830012<=p50 & group==1
_pctile rand if y0830013!=. & rand<. & group==1, nq(30)
scalar cutoff=r(r20)
replace y0830013=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
*Dropout pattern 14: (MAR 08 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 02 - higher dropout in group 1)
```

```
Deletion restricted to subjects with high last follow-up score in control group and low score in
experimental group
cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 5%/10%; visit 2 - 15%/25%; visit 3 - 20%/40% */
*************************************************************************************************************/
gen y0830021=y1
gen y0830022=y2
gen y0830023=y3

**group 0**
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r45)
replace y0830021=. if rand >=cutoff & rand<. & group==0
replace y0830022=. if y0830021==.
replace y0830023=. if y0830021==.
drop rand
_pctile y0830021 if group==0, nq(95)
scalar p50=r(r50)
gen rand=runiform() if y0830022!=. & y0830021>=p50 & group==0
_pctile rand if y0830022!=. & rand<. & group==0 , nq(45)
scalar cutoff=r(r35)
replace y0830022=. if rand >=cutoff & rand<.& group==0
replace y0830023=. if y0830022==.
drop rand
_pctile y0830022 if group==0, nq(85)
scalar p50=r(r50)
gen rand=runiform() if y0830023!=. & y0830022>=p50 & group==0
_pctile rand if y0830023!=. & rand<. & group==0, nq(35)
scalar cutoff=r(r30)
replace y0830023=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0<=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y0830021=. if rand >=cutoff & rand<. & group==1
replace y0830022=. if y0830021==.
replace y0830023=. if y0830021==.
drop rand
_pctile y0830021 if group==1, nq(90)
scalar p50=r(r40)
gen rand=runiform() if y0830022!=. & y0830021<=p50 & group==1
_pctile rand if y0830022!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r25)
replace y0830022=. if rand >=cutoff & rand<.& group==1
replace y0830023=. if y0830022==.
drop rand
_pctile y0830022 if group==1, nq(75)
scalar p50=r(r25)
gen rand=runiform() if y0830023!=. & y0830022<=p50 & group==1
_pctile rand if y0830023!=. & rand<. & group==1, nq(25)
scalar cutoff=r(r10)
replace y0830023=. if rand >=cutoff & rand<. & group==1
drop rand
```

```
/*****************************************************************************************************
Dropout pattern 15: (MAR 08 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 03 - higher dropout in group 0)
Deletion restricted to subjects with high last follow-up score in control group and low score in
experimental group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/5%; visit 2 - 25%/15%; visit 3 - 40%/20%
*****************************************************************************************************/
gen y0830031=y1
gen y0830032=y2
gen y0830033=y3

**group 0**
_pctile y0 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y0830031=. if rand >=cutoff & rand<. & group==0
replace y0830032=. if y0830031==.
replace y0830033=. if y0830031==.
drop rand
_pctile y0830031 if group==0, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y0830032!=. & y0830031>=p50 & group==0
_pctile rand if y0830032!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r25)
replace y0830032=. if rand >=cutoff & rand<.& group==0
replace y0830033=. if y0830032==.
drop rand
_pctile y0830032 if group==0, nq(75)
scalar p50=r(r50)
gen rand=runiform() if y0830033!=. & y0830032>=p50 & group==0
_pctile rand if y0830033!=. & rand<. & group==0, nq(25)
scalar cutoff=r(r10)
replace y0830033=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y0 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y0<=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r45)
replace y0830031=. if rand >=cutoff & rand<. & group==1
replace y0830032=. if y0830031==.
replace y0830033=. if y0830031==.
drop rand
_pctile y0830031 if group==1, nq(95)
scalar p50=r(r45)
gen rand=runiform() if y0830032!=. & y0830031<=p50 & group==1
_pctile rand if y0830032!=. & rand<. & group==1 , nq(45)
scalar cutoff=r(r35)
replace y0830032=. if rand >=cutoff & rand<.& group==1
replace y0830033=. if y0830032==.
drop rand

_pctile y0830032 if group==1, nq(85)
scalar p50=r(r35)
```

```
gen rand=runiform() if y0830033!=. & y0830032<=p50 & group==1
_pctile rand if y0830033!=. & rand<. & group==1, nq(35)
scalar cutoff=r(r30)
replace y0830033=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
Dropout pattern 16: (MNAR 10 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 01 - equal dropout between groups)
Deletion restricted to subjects with high score in each group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/10%; visit 2 - 20%/20%; visit 3 - 30%/30%
*****************************************************************************************************/
gen y1030011=y1
gen y1030012=y2
gen y1030013=y3

**group 0**
_pctile y1030011 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1030011>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y1030011=. if rand >=cutoff & rand<. & group==0
replace y1030012=. if y1030011==.
replace y1030013=. if y1030011==.
drop rand
_pctile y1030012 if group==0, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y1030012!=. & y1030012>=p50 & group==0
_pctile rand if y1030012!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r30)
replace y1030012=. if rand >=cutoff & rand<.& group==0
replace y1030013=. if y1030012==.
drop rand
_pctile y1030013 if group==0, nq(80)
scalar p50=r(r50)
gen rand=runiform() if y1030013!=. & y1030013>=p50 & group==0
_pctile rand if y1030013!=. & rand<. & group==0, nq(30)
scalar cutoff=r(r20)
replace y1030013=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y1030011 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1030011>=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y1030011=. if rand >=cutoff & rand<. & group==1
replace y1030012=. if y1030011==.
replace y1030013=. if y1030011==.
drop rand
_pctile y1030012 if group==1, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y1030012!=. & y1030012>=p50 & group==1
_pctile rand if y1030012!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r30)
replace y1030012=. if rand >=cutoff & rand<.& group==1
replace y1030013=. if y1030012==.
```

```
drop rand
_pctile y1030013 if group==1, nq(80)
scalar p50=r(r50)
gen rand=runiform() if y1030013!=. & y1030013>=p50 & group==1
_pctile rand if y1030013!=. & rand<. & group==1, nq(30)
scalar cutoff=r(r20)
replace y1030013=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
Dropout pattern 17: (MNAR 10 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 02 - higher dropout in group 1)
Deletion restricted to subjects with high score in each group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 5%/10%; visit 2 - 15%/25%; visit 3 - 20%/40%
*****************************************************************************************************/
gen y1030021=y1
gen y1030022=y2
gen y1030023=y3

**group 0**
_pctile y1030021 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1030021>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r45)
replace y1030021=. if rand >=cutoff & rand<. & group==0
replace y1030022=. if y1030021==.
replace y1030023=. if y1030021==.
drop rand
_pctile y1030022 if group==0, nq(95)
scalar p50=r(r50)
gen rand=runiform() if y1030022!=. & y1030022>=p50 & group==0
_pctile rand if y1030022!=. & rand<. & group==0 , nq(45)
scalar cutoff=r(r35)
replace y1030022=. if rand >=cutoff & rand<.& group==0
replace y1030023=. if y1030022==.
drop rand
_pctile y1030023 if group==0, nq(85)
scalar p50=r(r50)
gen rand=runiform() if y1030023!=. & y1030023>=p50 & group==0
_pctile rand if y1030023!=. & rand<. & group==0, nq(35)
scalar cutoff=r(r30)
replace y1030023=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y1030021 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1030021>=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y1030021=. if rand >=cutoff & rand<. & group==1
replace y1030022=. if y1030021==.
replace y1030023=. if y1030021==.
drop rand
_pctile y1030022 if group==1, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y1030022!=. & y1030022>=p50 & group==1
_pctile rand if y1030022!=. & rand<. & group==1 , nq(40)
```

```
scalar cutoff=r(r25)
replace y1030022=. if rand >=cutoff & rand<.& group==1
replace y1030023=. if y1030022==.
drop rand
_pctile y1030023 if group==1, nq(75)
scalar p50=r(r50)
gen rand=runiform() if y1030023!=. & y1030023>=p50 & group==1
_pctile rand if y1030023!=. & rand<. & group==1, nq(25)
scalar cutoff=r(r10)
replace y1030023=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
Dropout pattern 18: (MNAR 10 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 03 - higher dropout in group 0)
Deletion restricted to subjects with high score in each group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/5%; visit 2 - 25%/15%; visit 3 - 40%/20%
*****************************************************************************************************/
gen y1030031=y1
gen y1030032=y2
gen y1030033=y3

**group 0**
_pctile y1030031 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1030031>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y1030031=. if rand >=cutoff & rand<. & group==0
replace y1030032=. if y1030031==.
replace y1030033=. if y1030031==.
drop rand
_pctile y1030032 if group==0, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y1030032!=. & y1030032>=p50 & group==0
_pctile rand if y1030032!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r25)
replace y1030032=. if rand >=cutoff & rand<.& group==0
replace y1030033=. if y1030032==.
drop rand
_pctile y1030033 if group==0, nq(75)
scalar p50=r(r50)
gen rand=runiform() if y1030033!=. & y1030033>=p50 & group==0
_pctile rand if y1030033!=. & rand<. & group==0, nq(25)
scalar cutoff=r(r10)
replace y1030033=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y1030031 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1030031>=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r45)
replace y1030031=. if rand >=cutoff & rand<. & group==1
replace y1030032=. if y1030031==.
replace y1030033=. if y1030031==.
drop rand
_pctile y1030032 if group==1, nq(95)
```

```
scalar p50=r(r50)
gen rand=runiform() if y1030032!=. & y1030032>=p50 & group==1
_pctile rand if y1030032!=. & rand<. & group==1 , nq(45)
scalar cutoff=r(r35)
replace y1030032=. if rand >=cutoff & rand<.& group==1
replace y1030033=. if y1030032==.
drop rand
_pctile y1030033 if group==1, nq(85)
scalar p50=r(r50)
gen rand=runiform() if y1030033!=. & y1030033>=p50 & group==1
_pctile rand if y1030033!=. & rand<. & group==1, nq(35)
scalar cutoff=r(r30)
replace y1030033=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
*Dropout pattern 19: (MNAR 12 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 01 - equal dropout between groups)*
Deletion restricted to subjects with high score in control group and low score in experimental group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/10%; visit 2 - 20%/20%; visit 3 - 30%/30%
*****************************************************************************************************/
gen y1230011=y1
gen y1230012=y2
gen y1230013=y3

**group 0**
_pctile y1230011 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1230011>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y1230011=. if rand >=cutoff & rand<. & group==0
replace y1230012=. if y1230011==.
replace y1230013=. if y1230011==.
drop rand
_pctile y1230012 if group==0, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y1230012!=. & y1230012>=p50 & group==0
_pctile rand if y1230012!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r30)
replace y1230012=. if rand >=cutoff & rand<.& group==0
replace y1230013=. if y1230012==.
drop rand
_pctile y1230013 if group==0, nq(80)
scalar p50=r(r50)
gen rand=runiform() if y1230013!=. & y1230013>=p50 & group==0
_pctile rand if y1230013!=. & rand<. & group==0, nq(30)
scalar cutoff=r(r20)
replace y1230013=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y1230011 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1230011<=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
replace y1230011=. if rand >=cutoff & rand<. & group==1
replace y1230012=. if y1230011==.
```

```
replace y1230013=. if y1230011==.
drop rand
_pctile y1230012 if group==1, nq(90)
scalar p50=r(r40)
gen rand=runiform() if y1230012!=. & y1230012<=p50 & group==1
_pctile rand if y1230012!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r30)
replace y1230012=. if rand >=cutoff & rand<.& group==1
replace y1230013=. if y1230012==.
drop rand
_pctile y1230013 if group==1, nq(80)
scalar p50=r(r30)
gen rand=runiform() if y1230013!=. & y1230013<=p50 & group==1
_pctile rand if y1230013!=. & rand<. & group==1, nq(30)
scalar cutoff=r(r20)
replace y1230013=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************************
Dropout pattern 20: (MNAR 12 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 02 - higher dropout in group 1)*
Deletion restricted to subjects with high score in control group and low score in experimental group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 5%/10%; visit 2 - 15%/25%; visit 3 - 20%/40%
*****************************************************************************************************************/
gen y1230021=y1
gen y1230022=y2
gen y1230023=y3
**group 0**
_pctile y1230021 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1230021>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r45)
replace y1230021=. if rand >=cutoff & rand<. & group==0
replace y1230022=. if y1230021==.
replace y1230023=. if y1230021==.
drop rand
_pctile y1230022 if group==0, nq(95)
scalar p50=r(r50)
gen rand=runiform() if y1230022!=. & y1230022>=p50 & group==0
_pctile rand if y1230022!=. & rand<. & group==0 , nq(45)
scalar cutoff=r(r35)
replace y1230022=. if rand >=cutoff & rand<.& group==0
replace y1230023=. if y1230022==.
drop rand
_pctile y1230023 if group==0, nq(85)
scalar p50=r(r50)
gen rand=runiform() if y1230023!=. & y1230023>=p50 & group==0
_pctile rand if y1230023!=. & rand<. & group==0, nq(35)
scalar cutoff=r(r30)
replace y1230023=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y1230021 if group==1, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1230021<=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r40)
```

```
replace y1230021=. if rand >=cutoff & rand<. & group==1
replace y1230022=. if y1230021==.
replace y1230023=. if y1230021==.
drop rand
_pctile y1230022 if group==1, nq(90)
scalar p50=r(r40)
gen rand=runiform() if y1230022!=. & y1230022<=p50 & group==1
_pctile rand if y1230022!=. & rand<. & group==1 , nq(40)
scalar cutoff=r(r25)
replace y1230022=. if rand >=cutoff & rand<.& group==1
replace y1230023=. if y1230022==.
drop rand
_pctile y1230023 if group==1, nq(75)
scalar p50=r(r25)
gen rand=runiform() if y1230023!=. & y1230023<=p50 & group==1
_pctile rand if y1230023!=. & rand<. & group==1, nq(25)
scalar cutoff=r(r10)
replace y1230023=. if rand >=cutoff & rand<. & group==1
drop rand

/*****************************************************************************************************
*Dropout pattern 21: (MNAR 12 - dropout depends on last observed values; Dropout rate 30%;
Dropout direction 03 - higher dropout in group 0)*
Deletion restricted to subjects with high score in control group and low score in experimental group
Cumulative dropout rate: visit 0 - 0%/0%; visit 1 - 10%/5%; visit 2 - 25%/15%; visit 3 - 40%/20%
*****************************************************************************************************/
gen y1230031=y1
gen y1230032=y2
gen y1230033=y3

**group 0**
_pctile y1230031 if group==0, nq(100)
scalar p50=r(r50)
gen rand=runiform() if y1230031>=p50 & group==0
_pctile rand if rand<. & group==0 , nq(50)
scalar cutoff=r(r40)
replace y1230031=. if rand >=cutoff & rand<. & group==0
replace y1230032=. if y1230031==.
replace y1230033=. if y1230031==.
drop rand
_pctile y1230032 if group==0, nq(90)
scalar p50=r(r50)
gen rand=runiform() if y1230032!=. & y1230032>=p50 & group==0
_pctile rand if y1230032!=. & rand<. & group==0 , nq(40)
scalar cutoff=r(r25)
replace y1230032=. if rand >=cutoff & rand<.& group==0
replace y1230033=. if y1230032==.
drop rand
_pctile y1230033 if group==0, nq(75)
scalar p50=r(r50)
gen rand=runiform() if y1230033!=. & y1230033>=p50 & group==0
_pctile rand if y1230033!=. & rand<. & group==0, nq(25)
scalar cutoff=r(r10)
replace y1230033=. if rand >=cutoff & rand<. & group==0
drop rand

**group 1**
_pctile y1230031 if group==1, nq(100)
scalar p50=r(r50)
```

```
gen rand=runiform() if y1230031<=p50 & group==1
_pctile rand if rand<. & group==1 , nq(50)
scalar cutoff=r(r45)
replace y1230031=. if rand >=cutoff & rand<. & group==1
replace y1230032=. if y1230031==.
replace y1230033=. if y1230031==.
drop rand
_pctile y1230032 if group==1, nq(95)
scalar p50=r(r45)
gen rand=runiform() if y1230032!=. & y1230032<=p50 & group==1
_pctile rand if y1230032!=. & rand<. & group==1 , nq(45)
scalar cutoff=r(r35)
replace y1230032=. if rand >=cutoff & rand<.& group==1
replace y1230033=. if y1230032==.
drop rand
_pctile y1230033 if group==1, nq(85)
scalar p50=r(r35)
gen rand=runiform() if y1230033!=. & y1230033<=p50 & group==1
_pctile rand if y1230033!=. & rand<. & group==1, nq(35)
scalar cutoff=r(r30)
replace y1230033=. if rand >=cutoff & rand<. & group==1
drop rand

tempfile missdata
save `missdata',replace /*saving the dataset*/

/*****************************************************************************************************************
ANALYSIS
*****************************************************************************************************************/
/*ANALYSIS: No dropout*/
keep id group baseline y0 y1 y2 y3
****ANCOVA
regress y3 i.group y0
        return scalar at3=_b[1.group]
        return scalar ase3=_se[1.group]
        return scalar adf3= e(df_r)
        return scalar ap3=(2*ttail(e(df_r), abs(_b[1.group]/_se[1.group]))) /* 'ttail' returns the
reverse cumulative (upper tail or survivor) Student's t-distribution; it returns the probability T > t.*/
****MMRM
reshape long y , i(id) j(time)
gen rand1=group*(time==1)
gen rand2=group*(time==2)
gen rand3=group*(time==3)
*xtmixed y i.time i.rand1 i.rand2 i.rand3 || id:, reml nocons res(uns, t(time)) /*MMRM - baseline as
an outcome*/
xtmixed y i.time##c.baseline i.rand1 i.rand2 i.rand3 if time>0 || id:, reml nocons res(uns, t(time))
/*MMRM - baseline as an covariate*/
        return scalar bt3=_b[1.rand3]
        return scalar se3=_se[1.rand3]
        return scalar df3= e(N)
        return scalar p3=(2*normal(-abs(_b[1.rand3]/_se[1.rand3])))
foreach dr in 30 {
foreach dm in 01 02 04 06 08 10 12 {
/*ANALYSIS without IMPUTATION*/
use `missdata', clear
keep id group baseline y0 y1 y2 y3 y`dm'`dr'010 y`dm'`dr'011 y`dm'`dr'012 y`dm'`dr'013
y`dm'`dr'020 y`dm'`dr'021 y`dm'`dr'022 y`dm'`dr'023 y`dm'`dr'030 y`dm'`dr'031 y`dm'`dr'032
y`dm'`dr'033
****ANCOVA without IMPUTATION
```

```
        foreach dp in  01 02 03 {
        regress y`dm'`dr'`dp'3 i.group y0
        return scalar at`dm'`dr'`dp'3=_b[1.group]
        return scalar ase`dm'`dr'`dp'3=_se[1.group]
        return scalar adf`dm'`dr'`dp'3= e(df_r)
        return scalar ap`dm'`dr'`dp'3= (2*ttail(e(df_r), abs(_b[1.group]/_se[1.group])))
        }

****MMRM without IMPUTATION
        reshape long y y`dm'`dr'01 y`dm'`dr'02 y`dm'`dr'03, i(id) j(time)
        gen rand1=group*(time==1)
        gen rand2=group*(time==2)
        gen rand3=group*(time==3)
        foreach dp in  01 02 03 {
        *xtmixed y`dm'`dr'`dp' i.time i.rand1 i.rand2 i.rand3 || id:, reml nocons res(uns, t(time))
/*MMRM - baseline as an outcome*/
        xtmixed y`dm'`dr'`dp' i.time##c.baseline i.rand1 i.rand2 i.rand3 if time>0 || id:, reml nocons
res(uns, t(time)) /*MMRM - baseline as an covariate*/
        return scalar bt`dm'`dr'`dp'3=_b[1.rand3]
        return scalar se`dm'`dr'`dp'3=_se[1.rand3]
        return scalar df`dm'`dr'`dp'3= e(N)
        return scalar p`dm'`dr'`dp'3=(2*normal(-abs(_b[1.rand3]/_se[1.rand3])))
        }
/*ANALYSIS with LOCF */
use `missdata', clear
keep id group baseline y0 y1 y2 y3 y`dm'`dr'010 y`dm'`dr'011 y`dm'`dr'012 y`dm'`dr'013
y`dm'`dr'020 y`dm'`dr'021 y`dm'`dr'022 y`dm'`dr'023 y`dm'`dr'030 y`dm'`dr'031 y`dm'`dr'032
y`dm'`dr'033
/*Imputing with LOCF*/
foreach dp in  01 02 03 {
replace y`dm'`dr'`dp'1=y0 if y`dm'`dr'`dp'1==.
replace y`dm'`dr'`dp'2=y`dm'`dr'`dp'1 if y`dm'`dr'`dp'2==.
replace y`dm'`dr'`dp'3=y`dm'`dr'`dp'2 if y`dm'`dr'`dp'3==.
}
****LOCF ANCOVA
foreach dp in  01 02 03 {
regress y`dm'`dr'`dp'3 i.group y0
return scalar l_at`dm'`dr'`dp'3=_b[1.group]
return scalar l_ase`dm'`dr'`dp'3=_se[1.group]
return scalar l_adf`dm'`dr'`dp'3= e(df_r)
return scalar l_ap`dm'`dr'`dp'3= (2*ttail(e(df_r), abs(_b[1.group]/_se[1.group])))
}
//ANALYSIS with MI */
 tempname A B C
foreach dp in 01 02 03 {
use `missdata', clear
keep id group baseline y`dm'`dr'`dp'0 y`dm'`dr'`dp'1 y`dm'`dr'`dp'2 y`dm'`dr'`dp'3
mi set flong
mi reg imp y`dm'`dr'`dp'1 y`dm'`dr'`dp'2 y`dm'`dr'`dp'3
mi impute monotone (reg) y`dm'`dr'`dp'1 y`dm'`dr'`dp'2 y`dm'`dr'`dp'3 = baseline group, add(`dr')
/*MI ANCOVA*/
mi estimate: reg y`dm'`dr'`dp'3 baseline i.group
matrix A = e(b_mi)
matrix B = e(V_mi)
matrix C = e(df_mi)
return scalar m_at`dm'`dr'`dp'3=A[1,3]
return scalar m_ase`dm'`dr'`dp'3=sqrt(B[3,3])
return scalar m_adf`dm'`dr'`dp'3=C[1,3]
return scalar m_ap`dm'`dr'`dp'3=(2*ttail( C[1,3], abs(A[1,3]/sqrt(B[3,3]))))
```

```
}
}
}
clear
end
```

Appendix 4: Simulation results (tables 1 and 2)

Tables 1 and 2 presents the simulation results for scenarios contrasting to the missing data mechanisms used in the main text. In these tables, the '*same direction of dropouts*' was defined such that dropouts were subjects who did well in both study groups and the '*opposite direction of dropouts*' was defined such that dropouts were subjects who did well in the control group and those who did poorly in the experimental group.

Table 1: Bias and RMSE (in bracket)

Mechanism	30% dropouts			
	CCA	MMRM	LOCF	MI
No dropout	0.09 (2.68)	0.09 (2.68)		
Equal dropout between groups				
MAR-B1	0.09 (3.18)	0.10 (3.13)	-1.80 (3.08)	0.11 (3.15)
MAR-B2	0.0 (3.43)	0.02 (3.36)	-2.93 (3.87)	0.02 (3.38)
MAR-L1	0.04 (3.18)	0.01 (3.14)	-1.91 (3.12)	0.03 (3.15)
MAR-L2	2.56 (4.16)	0.10 (3.25)	-4.21 (4.88)	0.11 (3.28)
MNAR-1	0.12 (3.06)	0.15 (3.01)	-1.76 (2.97)	0.15 (3.04)
MNAR-2	5.57 (6.42)	4.28 (5.33)	-0.10 (2.44)	4.28 (5.33)
Higher dropout in the experimental group				
MAR-B1	0.01 (3.28)	0.05 (3.18)	-3.77 (4.52)	0.05 (3.20)
MAR-B2	0.16 (3.49)	0.12 (3.42)	-5.05 (5.67)	0.12 (3.43)
MAR-L1	-1.46 (3.58)	0.05 (3.23)	-3.19 (4.07)	0.04 (3.24)
MAR-L2	3.09 (4.59)	0.16 (3.37)	-6.45 (6.92)	0.16 (3.40)
MNAR-1	-3.23 (4.46)	-2.44 (3.93)	-5.02 (5.61)	-2.44 (3.94)
MNAR-2	6.43 (7.19)	4.92 (5.88)	-2.15 (3.29)	4.90 (5.87)
Higher dropout in the control group				
MAR-B1	0.02 (3.29)	0.0 (3.19)	0.24 (2.52)	-0.01 (3.19)
MAR-B2	0.17 (3.65)	0.17 (3.50)	-0.35 (2.56)	0.18 (3.53)
MAR-L1	1.62 (3.62)	0.08 (3.17)	-0.37 (2.49)	0.09 (3.18)
MAR-L2	2.98 (4.50)	0.06 (3.31)	-1.85 (3.09)	0.06 (3.33)
MNAR-1	3.41 (4.63)	2.60 (4.02)	1.53 (2.85)	2.60 (4.03)
MNAR-2	6.46 (7.21)	4.89 (5.85)	2.47 (3.52)	4.89 (5.87)

Table 2: Coverage of 95% CI and statistical power (in bracket)

Mechanism	30% dropouts			
	CCA	MMRM	LOCF	MI
No dropout	95.9 (91.5)	95.9 (91.5)		
Equal dropout between groups				
MAR-B1	96.2 (78.9)	95.8 (80.9)	91.8 (78.7)	96.0 (78.6)
MAR-B2	95.8 (72.7)	95.7 (75.2)	83.6 (60.8)	96.1 (73.0)
MAR-L1	95.0 (78.8)	95.2 (79.5)	88.8 (82.2)	95.9 (78.8)
MAR-L2	88.9 (93.0)	94.9 (79.8)	71.9 (37.9)	94.8 (75.9)
MNAR-1	95.6 (81.7)	95.8 (83.9)	89.8 (82.8)	95.9 (82.0)
MNAR-2	61.7 (99.2)	73.7 (98.4)	96.7 (91.6)	74.9 (98.0)
Higher dropout in the experimental group				
MAR-B1	96.2 (76.1)	96.2 (79.5)	72.1 (50.3)	96.9 (77.2)
MAR-B2	95.5 (73.1)	95.2 (74.3)	56.0 (28.2)	95.4 (72.1)
MAR-L1	93.0 (62.8)	94.9 (79.7)	75.7 (63.9)	95.1 (77.7)
MAR-L2	86.2 (94.6)	94.7 (77.7)	38.3 (10.5)	95.7 (75.0)
MNAR-1	84.5 (43.9)	87.9 (54.9)	47.9 (35.3)	88.9 (52.5)
MNAR-2	50.5 (99.7)	66.0 (98.8)	91.4 (68.9)	68.0 (98.6)
Higher dropout in the control group				
MAR-B1	95.0 (76.8)	94.9 (79.6)	96.1 (94.9)	95.8 (76.1)
MAR-B2	94.9 (73.1)	94.9 (77.5)	96.8 (90.2)	94.8 (74.7)
MAR-L1	93.4 (88.8)	95.8 (81.1)	94.5 (93.3)	96.4 (78.3)
MAR-L2	87.1 (92.6)	94.3 (78.3)	91.7 (76.3)	95.3 (75.0)
MNAR-1	82.7 (97.6)	86.7 (96.6)	91.5 (98.7)	88.1 (95.7)
MNAR-2	50.1 (99.6)	66.9 (98.6)	86.2 (99.3)	69.7 (98.2)

Appendix 5: Simulation results (tables 3–22)

Tables 3–22 present the simulation results, which presented graphically in chapter 5, in tabular form.

Table 3: Bias under MCAR

Dropout rate		Variance-covariance	CCA	MMRM	LOCF	MI
%	between groups					
10%	EQ	WL	0.04	0.05	−0.74	0.06
		WM	−0.03	−0.04	−0.69	−0.05
		WH	−0.01	−0.01	−0.62	0.01
		SL	−0.10	−0.14	−1.09	−0.13
		SM	−0.06	−0.05	−0.85	−0.05
		SH	0.07	0.11	−0.54	0.13
	HE	WL	0.11	0.11	−2.00	0.09
		WM	−0.05	−0.05	−1.46	−0.04
		WH	−0.05	−0.04	−1.55	−0.04
		SL	−0.11	−0.12	−2.36	−0.12
		SM	−0.06	−0.07	−1.67	−0.08
		SH	0.11	0.10	−1.62	0.10
	HC	WL	0.22	0.20	0.61	0.20
		WM	−0.04	−0.03	0.03	−0.04
		WH	0.03	0.03	0.23	0.04
		SL	−0.11	−0.11	0.24	−0.11
		SM	−0.06	−0.06	−0.10	−0.04
		SH	0.08	0.09	0.41	0.10
30%	EQ	WL	0.08	0.12	−1.92	0.12
		WM	0.02	0.04	−2.00	0.03
		WH	−0.11	−0.09	−1.98	−0.09
		SL	−0.01	−0.07	−1.97	−0.08
		SM	−0.01	0.02	−1.89	0.03
		SH	0.13	0.19	−1.75	0.20
	HE	WL	−0.05	−0.02	−4.52	−0.01
		WM	−0.02	−0.04	−4.36	−0.05
		WH	−0.19	−0.18	−4.24	−0.18
		SL	−0.13	−0.15	−4.39	−0.16
		SM	0.04	0.01	−4.17	−0.01
		SH	0.12	0.15	−4.05	0.14
	HC	WL	0.12	0.10	0.59	0.10
		WM	−0.01	−0.02	0.38	−0.02
		WH	0.02	0.03	0.43	0.01
		SL	−0.06	−0.09	0.20	−0.09
		SM	−0.03	−0.05	0.45	−0.04
		SH	0.10	0.09	0.49	0.11

Table 4: RMSE under MCAR

Dropout rate		Variance-covariance	CCA	MMRM	LOCF	MI
%	between groups					
10%	EQ	WL	2.89	2.87	2.73	2.89
		WM	2.91	2.90	2.76	2.93
		WH	2.83	2.82	2.71	2.85
		SL	2.96	2.91	2.79	2.92
		SM	2.96	2.94	2.81	2.96
		SH	3.02	2.99	2.80	3.00
	HE	WL	2.90	2.87	3.27	2.90
		WM	2.98	2.98	3.11	2.99
		WH	2.89	2.88	3.11	2.90
		SL	2.91	2.85	3.46	2.88
		SM	2.93	2.90	3.10	2.91
		SH	3.01	2.97	3.17	3.00
	HC	WL	2.86	2.85	2.72	2.87
		WM	3.00	2.99	2.74	3.02
		WH	2.92	2.91	2.70	2.93
		SL	2.94	2.91	2.59	2.95
		SM	2.96	2.92	2.65	2.95
		SH	2.95	2.94	2.75	2.96
30%	EQ	WL	3.36	3.34	3.22	3.37
		WM	3.37	3.35	3.30	3.35
		WH	3.28	3.23	3.28	3.22
		SL	3.27	3.18	3.16	3.21
		SM	3.25	3.22	3.15	3.23
		SH	3.28	3.21	3.09	3.23
	HE	WL	3.33	3.27	5.16	3.28
		WM	3.37	3.34	5.06	3.35
		WH	3.29	3.29	4.99	3.30
		SL	3.45	3.36	5.06	3.37
		SM	3.28	3.24	4.88	3.26
		SH	3.39	3.29	4.80	3.31
	HC	WL	3.43	3.35	2.60	3.38
		WM	3.40	3.35	2.58	3.38
		WH	3.24	3.21	2.60	3.22
		SL	3.42	3.30	2.49	3.35
		SM	3.24	3.19	2.56	3.20
		SH	3.37	3.27	2.59	3.27

Table 5: Bias under MAR-B

Dropout rate		Variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MAR-B1	MAR-B2	MAR-B1	MAR-B2	MAR-B1	MAR-B2	MAR-B1	MAR-B2
10%	EQ	WL	0.16	0.16	0.14	0.12	−0.70	−0.10	0.16	0.12
		WM	0.04	0.02	0.04	0.04	−0.62	0.14	0.04	0.04
		WH	−0.03	−0.03	−0.02	−0.03	−0.63	0.74	−0.01	−0.04
		SL	−0.10	−0.06	−0.10	−0.07	−0.86	−0.37	−0.13	−0.07
		SM	−0.04	−0.04	−0.03	−0.03	−0.82	−0.21	−0.05	−0.04
		SH	0.11	0.11	0.08	0.10	−0.57	0.66	0.09	0.10
	HE	WL	0.11	0.08	0.11	0.09	−2.24	−1.23	0.10	0.11
		WM	−0.01	−0.07	0.00	−0.06	−2.06	−0.73	0.00	−0.07
		WH	−0.01	0.01	−0.01	0.01	−2.06	0.15	−0.01	0.03
		SL	−0.03	−0.13	−0.03	−0.11	−2.43	−1.68	−0.02	−0.11
		SM	−0.02	−0.10	−0.01	−0.12	−1.77	−1.01	−0.01	−0.11
		SH	0.09	0.13	0.11	0.12	−2.06	−0.25	0.11	0.12
	HC	WL	0.17	0.15	0.16	0.14	0.84	1.33	0.17	0.12
		WM	0.02	−0.02	0.01	0.00	0.54	1.22	0.02	−0.01
		WH	0.00	−0.03	0.01	−0.02	0.70	1.89	0.01	0.00
		SL	−0.13	−0.08	−0.12	−0.10	0.51	0.90	−0.12	−0.11
		SM	−0.07	−0.04	−0.06	−0.03	0.06	0.60	−0.06	−0.02
		SH	0.12	0.07	0.10	0.08	0.89	1.77	0.10	0.09
30%	EQ	WL	0.09	0.12	0.07	0.13	−1.96	−0.57	0.06	0.14
		WM	0.04	−0.16	0.04	−0.16	−2.01	−0.30	0.04	−0.15
		WH	0.04	0.06	0.03	0.06	−1.90	1.65	0.03	0.05
		SL	−0.06	−0.11	−0.08	−0.12	−1.91	−1.11	−0.09	−0.11
		SM	0.01	−0.03	−0.01	0.01	−1.93	−0.80	−0.02	0.02
		SH	0.08	0.16	0.05	0.16	−1.85	0.85	0.06	0.16
	HE	WL	0.13	0.06	0.13	0.02	−4.76	−3.32	0.12	0.01
		WM	0.02	−0.25	0.02	−0.20	−4.78	−2.81	0.00	−0.20
		WH	−0.01	−0.18	0.01	−0.14	−5.08	−1.17	0.01	−0.13
		SL	−0.01	−0.10	0.00	−0.15	−4.44	−3.52	0.00	−0.16
		SM	−0.01	−0.05	−0.06	−0.04	−4.55	−3.20	−0.05	−0.04
		SH	0.03	0.15	0.02	0.15	−4.86	−1.65	0.03	0.15
	HC	WL	0.12	0.09	0.14	0.12	1.05	1.83	0.11	0.11
		WM	−0.06	−0.05	−0.05	−0.01	0.90	1.94	−0.05	0.00
		WH	−0.09	−0.13	−0.08	−0.14	1.26	3.38	−0.08	−0.12
		SL	−0.15	−0.09	−0.13	−0.09	0.48	1.07	−0.14	−0.10
		SM	0.05	0.07	0.05	0.11	0.84	1.49	0.06	0.12
		SH	0.05	0.10	0.03	0.11	1.17	2.98	0.03	0.11

Table 6: RMSE under MAR-B

Dropout rate		Variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MAR-B1	MAR-B2	MAR-B1	MAR-B2	MAR-B1	MAR-B2	MAR-B1	MAR-B2
10%	EQ	WL	2.84	2.87	2.84	2.87	2.75	2.64	2.88	2.89
		WM	2.90	2.92	2.90	2.89	2.79	2.64	2.91	2.91
		WH	2.85	2.90	2.84	2.88	2.73	2.73	2.85	2.89
		SL	2.89	2.90	2.87	2.86	2.78	2.57	2.88	2.87
		SM	2.94	3.02	2.90	2.96	2.78	2.66	2.93	2.97
		SH	2.97	3.01	2.95	2.97	2.81	2.79	2.95	2.98
	HE	WL	2.85	2.95	2.83	2.91	3.43	2.87	2.83	2.94
		WM	2.96	3.07	2.96	3.05	3.42	2.82	2.99	3.06
		WH	2.86	2.92	2.86	2.90	3.37	2.65	2.88	2.95
		SL	2.93	2.96	2.91	2.92	3.61	3.07	2.93	2.96
		SM	2.97	3.04	2.95	3.00	3.21	2.85	2.96	3.02
		SH	2.98	3.02	2.95	2.99	3.45	2.72	2.97	3.02
	HC	WL	2.84	2.86	2.83	2.86	2.75	2.94	2.84	2.90
		WM	2.92	3.01	2.92	3.00	2.74	2.95	2.95	3.03
		WH	2.91	2.94	2.88	2.93	2.72	3.29	2.90	2.92
		SL	2.95	3.02	2.90	2.99	2.66	2.74	2.92	2.99
		SM	2.97	3.02	2.94	2.98	2.70	2.72	2.96	3.01
		SH	2.97	2.96	2.93	2.94	2.82	3.22	2.94	2.96
30%	EQ	WL	3.28	3.62	3.25	3.56	3.23	2.64	3.25	3.61
		WM	3.30	3.60	3.28	3.56	3.27	2.53	3.30	3.55
		WH	3.20	3.47	3.16	3.40	3.14	3.02	3.19	3.42
		SL	3.19	3.38	3.11	3.27	3.18	2.67	3.11	3.28
		SM	3.29	3.56	3.22	3.45	3.24	2.65	3.23	3.46
		SH	3.31	3.67	3.25	3.52	3.17	2.67	3.27	3.54
	HE	WL	3.36	3.59	3.32	3.55	5.41	4.17	3.35	3.59
		WM	3.31	3.81	3.28	3.75	5.43	3.83	3.30	3.75
		WH	3.45	3.58	3.38	3.54	5.68	2.91	3.40	3.55
		SL	3.44	3.73	3.34	3.58	5.15	4.30	3.37	3.60
		SM	3.36	3.61	3.27	3.51	5.22	4.08	3.31	3.55
		SH	3.38	3.74	3.33	3.60	5.53	3.08	3.33	3.63
	HC	WL	3.32	3.53	3.31	3.46	2.84	3.12	3.33	3.47
		WM	3.35	3.55	3.30	3.49	2.70	3.15	3.33	3.51
		WH	3.32	3.62	3.29	3.58	2.86	4.29	3.31	3.59
		SL	3.43	3.81	3.30	3.62	2.62	2.68	3.32	3.64
		SM	3.33	3.52	3.26	3.38	2.73	2.91	3.26	3.38
		SH	3.43	3.71	3.36	3.55	2.86	3.95	3.38	3.56

Table 7: Bias under MAR-L

Dropout rate		variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MAR-L1	MAR-L2	MAR-L1	MAR-L2	MAR-L1	MAR-L2	MAR-L1	MAR-L2
10%	EQ	WL	0.08	−0.71	0.08	0.14	−0.71	1.35	0.08	0.14
		WM	0.00	−0.99	0.01	−0.06	−0.62	1.60	0.01	−0.04
		WH	−0.13	−1.56	−0.13	−0.09	−0.76	3.33	−0.13	−0.09
		SL	−0.10	−1.29	−0.09	−0.10	−1.00	0.28	−0.11	−0.10
		SM	0.00	−1.27	0.00	0.00	−0.81	0.60	−0.01	0.01
		SH	0.16	−1.64	0.14	0.11	−0.54	1.91	0.15	0.11
	HE	WL	0.49	−0.64	0.08	0.17	−2.85	0.09	0.08	0.17
		WM	0.42	−1.11	−0.06	−0.05	−3.02	0.72	−0.06	−0.05
		WH	0.59	−1.48	−0.07	−0.02	−3.12	2.62	−0.07	−0.01
		SL	0.27	−0.99	−0.09	−0.08	−2.68	−1.11	−0.08	−0.09
		SM	0.40	−1.33	−0.04	−0.08	−2.04	−0.23	−0.03	−0.08
		SH	0.84	−1.43	0.11	0.12	−2.54	0.83	0.11	0.11
	HC	WL	−0.27	−0.64	0.13	0.16	1.39	2.62	0.12	0.16
		WM	−0.50	−1.08	−0.02	−0.01	1.37	3.02	−0.03	−0.01
		WH	−0.73	−1.54	−0.08	−0.06	1.66	4.35	−0.08	−0.05
		SL	−0.40	−0.98	−0.04	−0.12	0.74	1.43	−0.03	−0.11
		SM	−0.45	−1.33	0.00	−0.08	0.34	1.34	0.01	−0.08
		SH	−0.63	−1.42	0.08	0.13	1.33	2.88	0.09	0.15
30%	EQ	WL	0.19	−1.50	0.18	0.21	−1.93	1.66	0.19	0.21
		WM	0.00	−2.51	0.01	−0.06	−2.05	2.92	0.01	−0.07
		WH	−0.11	−3.61	−0.10	−0.12	−1.98	6.44	−0.09	−0.14
		SL	−0.09	−1.81	−0.09	−0.16	−1.92	−0.21	−0.07	−0.17
		SM	0.00	−2.59	−0.02	−0.11	−1.92	0.41	−0.02	−0.11
		SH	0.14	−3.62	0.11	0.09	−1.80	2.93	0.11	0.09
	HE	WL	1.26	−1.94	0.20	0.09	−5.72	−0.76	0.22	0.09
		WM	1.39	−2.96	−0.09	−0.01	−6.44	0.64	−0.09	−0.02
		WH	2.05	−4.36	−0.07	−0.11	−7.39	4.43	−0.07	−0.09
		SL	0.94	−2.50	−0.09	−0.03	−4.49	−1.92	−0.06	0.00
		SM	1.47	−2.93	−0.08	0.01	−5.16	−1.74	−0.08	0.01
		SH	2.40	−4.50	0.08	0.17	−5.70	0.93	0.10	0.17
	HC	WL	−1.05	−1.91	−0.03	0.13	1.91	4.32	−0.04	0.13
		WM	−1.48	−2.93	0.00	−0.01	2.41	5.59	−0.01	0.02
		WH	−2.30	−4.36	−0.15	−0.12	3.48	9.07	−0.16	−0.14
		SL	−1.15	−2.57	−0.16	−0.22	0.15	1.65	−0.17	−0.21
		SM	−1.53	−2.86	0.00	0.07	1.40	2.88	0.01	0.06
		SH	−2.27	−4.54	0.02	0.11	2.01	5.36	0.02	0.14

Table 8: RMSE under MAR-L

Dropout rate		variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MAR-L1	MAR-L2	MAR-L1	MAR-L2	MAR-L1	MAR-L2	MAR-L1	MAR-L2
10%	EQ	WL	2.82	2.98	2.82	2.89	2.67	2.92	2.83	2.90
		WM	2.90	3.09	2.89	2.93	2.72	3.11	2.89	2.95
		WH	2.87	3.31	2.86	2.92	2.73	4.26	2.90	2.93
		SL	2.92	3.24	2.91	2.94	2.79	2.54	2.94	2.98
		SM	2.88	3.26	2.87	3.00	2.75	2.74	2.91	3.01
		SH	3.01	3.42	2.97	3.00	2.78	3.34	2.99	3.03
	HE	WL	2.89	3.00	2.84	2.93	3.82	2.59	2.86	2.98
		WM	2.98	3.16	2.95	2.96	4.02	2.74	2.97	2.97
		WH	2.96	3.23	2.89	2.87	4.10	3.70	2.91	2.89
		SL	2.92	3.14	2.89	2.94	3.73	2.78	2.90	2.96
		SM	3.00	3.27	2.98	2.95	3.40	2.62	2.97	2.95
		SH	3.13	3.28	3.00	2.95	3.76	2.81	3.00	2.96
	HC	WL	2.89	2.92	2.86	2.86	2.93	3.65	2.88	2.88
		WM	2.98	3.12	2.93	2.93	2.95	4.00	2.93	2.95
		WH	2.99	3.24	2.89	2.86	3.11	5.07	2.93	2.90
		SL	2.94	3.08	2.90	2.91	2.68	2.91	2.91	2.93
		SM	3.03	3.33	2.99	3.05	2.73	3.03	3.00	3.06
		SH	3.05	3.32	2.96	3.01	3.01	3.96	2.97	3.03
30%	EQ	WL	3.25	3.80	3.24	3.52	3.14	3.02	3.25	3.55
		WM	3.21	4.11	3.20	3.30	3.19	3.79	3.22	3.32
		WH	3.22	4.95	3.20	3.43	3.15	6.92	3.21	3.42
		SL	3.22	3.76	3.19	3.27	3.19	2.41	3.22	3.28
		SM	3.33	4.30	3.30	3.42	3.22	2.52	3.31	3.45
		SH	3.35	4.99	3.31	3.42	3.16	3.86	3.34	3.43
	HE	WL	3.51	3.98	3.31	3.51	6.24	2.59	3.35	3.53
		WM	3.73	4.58	3.45	3.60	6.91	2.62	3.48	3.61
		WH	3.90	5.56	3.31	3.52	7.80	5.11	3.32	3.55
		SL	3.45	4.25	3.28	3.47	5.16	3.08	3.32	3.49
		SM	3.63	4.59	3.29	3.50	5.75	3.04	3.31	3.53
		SH	4.22	5.67	3.44	3.42	6.28	2.67	3.46	3.45
	HC	WL	3.45	3.87	3.29	3.40	3.17	4.96	3.32	3.41
		WM	3.64	4.51	3.30	3.48	3.46	6.14	3.32	3.51
		WH	4.03	5.50	3.31	3.38	4.29	9.40	3.34	3.40
		SL	3.59	4.36	3.41	3.52	2.54	2.92	3.43	3.53
		SM	3.64	4.42	3.29	3.36	2.92	3.80	3.31	3.37
		SH	4.07	5.72	3.36	3.45	3.27	5.92	3.39	3.49

Table 9(A): Bias under MNAR in relation to dropout rate

variance-covariance	Dropout rate		CCA		MMRM		LOCF		MI	
	between groups	%	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2
SM	EQ	10%	0.00	−2.77	−0.01	−2.11	−0.81	−1.79	0.00	−2.13
		20%	0.06	−4.70	0.04	−3.56	−1.22	−2.91	0.03	−3.56
		30%	0.03	−5.56	0.05	−4.23	−1.84	−3.62	0.06	−4.24
		40%	0.19	−8.17	0.15	−6.20	−2.18	−4.78	0.16	−6.20
		50%	0.07	−9.41	0.04	−6.94	−2.75	−5.31	0.05	−6.94
	HE	10%	0.99	−2.84	0.76	−2.20	−1.20	−2.62	0.77	−2.19
		20%	1.72	−4.90	1.33	−3.76	−2.20	−4.38	1.32	−3.74
		30%	3.41	−6.42	2.62	−4.92	−3.10	−6.11	2.61	−4.92
		40%	3.80	−8.12	2.91	−6.17	−3.83	−7.00	2.90	−6.17
		50%	4.51	−10.16	3.39	−7.59	−4.16	−7.40	3.41	−7.59
	HC	10%	−0.93	−2.85	−0.72	−2.21	−0.39	−1.07	−0.71	−2.22
		20%	−1.70	−4.90	−1.28	−3.77	−0.43	−1.61	−1.29	−3.78
		30%	−3.24	−6.26	−2.43	−4.73	−0.42	−1.37	−2.43	−4.74
		40%	−3.72	−8.15	−2.82	−6.16	−1.03	−2.37	−2.84	−6.17
		50%	−4.62	−10.18	−3.49	−7.62	−1.71	−3.33	−3.49	−7.61

Table 9(B): Bias under MNAR in relation to covariance pattern

Dropout rate		variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2
30%	EQ	WL	0.14	−4.69	0.14	−4.11	−1.97	−3.92	0.14	−4.08
		WM	−0.04	−6.90	−0.04	−6.06	−2.06	−4.98	−0.03	−6.07
		WH	−0.01	−9.79	−0.01	−8.58	−1.91	−5.63	−0.02	−8.59
		SL	−0.15	−3.86	−0.20	−2.93	−2.03	−2.98	−0.20	−2.94
		SM	−0.20	−5.68	−0.20	−4.34	−2.08	−3.67	−0.18	−4.35
		SH	0.07	−8.37	0.04	−6.36	−1.88	−3.62	0.04	−6.37
	HE	WL	3.18	−5.26	2.81	−4.59	−3.22	−6.46	2.81	−4.60
		WM	4.35	−7.86	3.86	−6.88	−2.71	−7.61	3.85	−6.87
		WH	6.21	−11.46	5.49	−10.05	−1.79	−8.46	5.48	−10.04
		SL	2.16	−5.24	1.62	−3.97	−3.34	−5.34	1.64	−3.96
		SM	3.25	−6.45	2.45	−4.89	−3.23	−6.12	2.44	−4.91
		SH	5.11	−9.94	3.93	−7.49	−2.70	−6.23	3.93	−7.48
	HC	WL	−2.92	−5.40	−2.54	−4.71	−0.53	−1.52	−2.53	−4.71
		WM	−4.44	−7.95	−3.94	−6.98	−1.26	−2.76	−3.92	−6.98
		WH	−6.35	−11.39	−5.58	−9.94	−1.98	−3.71	−5.57	−9.95
		SL	−2.31	−5.18	−1.75	−3.91	−0.89	−1.65	−1.74	−3.90
		SM	−3.29	−6.41	−2.45	−4.85	−0.46	−1.45	−2.44	−4.84
		SH	−4.87	−9.91	−3.70	−7.47	−0.84	−1.74	−3.70	−7.46

Table 10(A): RMSE under MNAR in relation to dropout rate

variance-covariance	Dropout rate		CCA		MMRM		LOCF		MI	
	between groups	%	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2
SM	EQ	10%	2.93	4.02	2.92	3.60	2.79	3.20	2.91	3.62
		20%	3.05	5.59	3.02	4.69	2.88	3.87	3.04	4.70
		30%	3.25	6.40	3.24	5.32	3.17	4.39	3.24	5.32
		40%	3.38	8.85	3.34	7.08	3.32	5.35	3.35	7.10
		50%	3.60	10.19	3.55	7.97	3.68	5.84	3.54	7.97
	HE	10%	3.06	4.05	2.98	3.65	2.89	3.71	3.00	3.65
		20%	3.53	5.77	3.36	4.84	3.46	5.06	3.37	4.85
		30%	4.69	7.21	4.14	5.93	4.02	6.62	4.15	5.92
		40%	5.11	8.82	4.49	7.10	4.57	7.41	4.50	7.11
		50%	5.89	10.90	5.02	8.54	4.79	7.75	5.05	8.56
	HC	10%	3.05	4.07	2.98	3.66	2.66	2.84	3.00	3.69
		20%	3.51	5.80	3.30	4.87	2.61	3.00	3.31	4.89
		30%	4.57	7.09	4.05	5.77	2.60	2.87	4.06	5.79
		40%	5.17	8.88	4.54	7.10	2.73	3.35	4.56	7.12
		50%	5.91	10.94	5.06	8.62	2.94	4.10	5.07	8.63

Table 10(B): RMSE under MNAR in relation to covariance pattern

Dropout rate		variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2
30%	EQ	WL	3.11	5.70	3.14	5.23	3.23	4.68	3.16	5.21
		WM	3.17	7.62	3.20	6.87	3.29	5.59	3.21	6.89
		WH	3.22	10.26	3.21	9.13	3.23	6.17	3.23	9.14
		SL	3.21	5.01	3.21	4.34	3.29	3.86	3.23	4.34
		SM	3.17	6.56	3.19	5.44	3.29	4.44	3.21	5.46
		SH	3.23	9.01	3.18	7.18	3.15	4.47	3.20	7.20
	HE	WL	4.43	6.19	4.19	5.63	4.06	6.94	4.20	5.66
		WM	5.41	8.48	5.02	7.58	3.73	8.01	5.02	7.58
		WH	6.97	11.91	6.34	10.56	3.17	8.86	6.34	10.55
		SL	3.89	6.20	3.66	5.16	4.21	5.87	3.68	5.19
		SM	4.59	7.22	4.04	5.85	4.05	6.60	4.04	5.88
		SH	6.10	10.46	5.11	8.17	3.73	6.74	5.10	8.17
	HC	WL	4.27	6.28	4.03	5.69	2.59	2.94	4.03	5.70
		WM	5.50	8.55	5.11	7.65	2.86	3.67	5.10	7.66
		WH	7.06	11.80	6.39	10.41	3.20	4.51	6.39	10.43
		SL	4.00	6.15	3.68	5.11	2.64	2.88	3.69	5.12
		SM	4.55	7.19	3.98	5.82	2.56	2.90	3.99	5.82
		SH	5.92	10.46	4.97	8.17	2.74	3.07	4.98	8.17

Table 11: Coverage under MCAR

Dropout rate		Variance-covariance	CCA	MMRM	LOCF	MI
%	between groups					
10%	EQ	WL	96.0%	95.5%	96.5%	95.7%
		WM	95.2%	95.3%	94.9%	94.8%
		WH	96.4%	96.4%	95.8%	95.8%
		SL	94.7%	94.9%	96.0%	95.6%
		SM	94.3%	94.4%	93.7%	94.1%
		SH	94.5%	94.5%	94.6%	94.0%
	HE	WL	95.2%	95.1%	92.5%	95.6%
		WM	94.5%	94.5%	92.1%	95.1%
		WH	95.7%	95.9%	93.2%	95.9%
		SL	94.8%	95.2%	90.6%	95.6%
		SM	95.4%	95.1%	92.3%	96.3%
		SH	94.7%	95.2%	90.8%	95.0%
	HC	WL	95.5%	95.5%	95.2%	96.2%
		WM	94.5%	94.0%	96.0%	94.7%
		WH	94.8%	94.8%	95.5%	94.6%
		SL	94.6%	94.7%	97.4%	95.2%
		SM	94.3%	94.8%	96.0%	95.3%
		SH	94.8%	94.3%	94.8%	94.7%
30%	EQ	WL	95.2%	94.6%	92.0%	95.6%
		WM	96.1%	95.5%	90.3%	96.0%
		WH	94.7%	94.9%	90.2%	95.5%
		SL	94.5%	94.8%	93.9%	95.2%
		SM	94.9%	94.7%	92.0%	95.5%
		SH	95.8%	95.1%	91.5%	95.2%
	HE	WL	94.7%	94.9%	69.4%	95.9%
		WM	95.6%	94.7%	66.3%	95.0%
		WH	95.4%	95.8%	66.0%	95.4%
		SL	95.1%	94.1%	71.8%	95.1%
		SM	94.9%	94.0%	69.3%	94.8%
		SH	94.6%	94.1%	67.4%	94.8%
	HC	WL	94.0%	93.7%	97.0%	94.7%
		WM	94.3%	95.6%	96.6%	95.5%
		WH	96.4%	96.3%	95.9%	96.0%
		SL	94.3%	94.2%	97.6%	95.4%
		SM	96.4%	95.7%	96.3%	96.0%
		SH	94.7%	95.1%	95.1%	95.3%

Table 12: Average width of CI under MCAR

Dropout rate		Variance-covariance	CCA	MMRM	LOCF	MI
%	between groups					
10%	EQ	WL	11.67	11.56	11.36	11.84
		WM	11.52	11.46	10.98	11.61
		WH	11.50	11.45	10.85	11.56
		SL	11.77	11.55	11.46	11.96
		SM	11.58	11.45	10.91	11.63
		SH	11.58	11.47	10.82	11.57
	HE	WL	11.69	11.58	11.50	11.88
		WM	11.58	11.51	11.01	11.65
		WH	11.53	11.48	10.85	11.58
		SL	11.68	11.50	11.66	11.87
		SM	11.63	11.49	10.97	11.69
		SH	11.57	11.47	10.85	11.59
	HC	WL	11.69	11.59	11.27	11.91
		WM	11.58	11.52	10.95	11.66
		WH	11.53	11.48	10.82	11.59
		SL	11.69	11.50	11.33	11.86
		SM	11.63	11.49	10.88	11.66
		SH	11.58	11.48	10.79	11.61
30%	EQ	WL	13.25	12.96	11.82	13.64
		WM	13.28	13.05	10.97	13.42
		WH	13.03	12.86	10.68	13.07
		SL	12.97	12.53	11.94	13.22
		SM	13.00	12.65	10.90	13.00
		SH	13.07	12.76	10.57	12.96
	HE	WL	13.47	13.13	11.84	13.81
		WM	13.35	13.11	11.02	13.47
		WH	13.22	13.02	10.73	13.26
		SL	13.45	12.86	11.91	13.67
		SM	13.11	12.70	10.91	13.08
		SH	13.29	12.91	10.62	13.14
	HC	WL	13.46	13.11	11.62	13.84
		WM	13.34	13.10	10.88	13.49
		WH	13.20	13.01	10.69	13.21
		SL	13.46	12.87	11.68	13.69
		SM	13.10	12.71	10.77	13.09
		SH	13.28	12.90	10.55	13.13

Table 11: Coverage (%) under MAR-B

Dropout rate		Variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MAR-B1	MAR-B2	MAR-B1	MAR-B2	MAR-B1	MAR-B2	MAR-B1	MAR-B2
10%	EQ	WL	95.1	95.4	94.8	95.1	96.2	96.3	95.4	96.0
		WM	95.0	95.1	94.9	95.0	95.4	96.5	94.6	95.5
		WH	95.6	96.9	95.6	96.4	95.1	95.0	95.5	96.3
		SL	94.7	94.6	94.9	94.3	95.5	96.6	95.3	95.1
		SM	94.6	94.9	95.1	95.6	95.1	95.7	95.2	95.1
		SH	95.0	94.0	94.7	93.7	94.7	94.6	95.1	94.5
	HE	WL	95.6	95.0	95.1	95.1	91.3	94.9	95.6	95.4
		WM	95.3	94.6	95.4	94.3	89.7	94.6	95.5	94.6
		WH	96.0	95.0	95.8	95.1	89.4	95.6	95.6	94.7
		SL	94.7	95.1	94.8	94.8	88.7	93.1	95.3	95.5
		SM	94.0	94.9	93.9	94.1	91.8	94.8	94.3	94.8
		SH	94.5	94.6	93.9	94.4	89.0	95.2	94.6	94.6
	HC	WL	95.3	95.5	95.1	95.3	95.6	93.2	95.6	95.8
		WM	95.2	94.2	94.9	94.9	96.2	92.9	95.4	94.6
		WH	95.3	95.0	95.2	94.9	94.9	89.6	95.0	95.8
		SL	95.0	94.5	94.9	94.4	97.2	96.0	95.7	95.5
		SM	95.0	95.1	94.6	95.3	96.1	95.1	95.0	95.1
		SH	95.3	95.0	95.1	95.4	94.2	91.2	95.2	95.2
30%	EQ	WL	94.6	95.1	94.8	95.1	93.9	98.0	95.4	96.0
		WM	95.5	95.8	94.7	94.9	91.0	97.5	95.0	95.4
		WH	95.1	95.6	95.3	95.2	91.1	92.1	95.3	95.7
		SL	95.0	95.8	94.5	95.2	92.5	97.6	96.0	95.4
		SM	95.5	93.3	95.4	94.4	91.8	95.8	95.7	94.4
		SH	94.5	94.0	94.5	94.7	91.6	95.9	95.0	94.7
	HE	WL	95.4	94.8	94.7	95.1	66.4	83.9	96.1	96.4
		WM	95.6	93.8	95.1	93.5	60.1	84.1	95.9	94.5
		WH	94.2	95.1	94.6	94.5	55.8	92.3	94.7	94.7
		SL	95.2	94.6	95.1	93.2	68.7	82.5	95.8	95.4
		SM	94.1	94.7	94.8	94.1	63.8	80.4	94.9	95.3
		SH	94.9	94.7	94.2	94.4	57.7	91.1	94.8	94.5
	HC	WL	95.1	96.6	95.2	95.8	95.4	92.9	95.7	96.5
		WM	95.4	94.8	95.3	95.5	96.1	91.5	95.7	95.3
		WH	95.5	93.4	95.1	93.5	94.5	77.8	95.3	93.8
		SL	94.8	94.8	94.6	94.1	96.4	96.8	95.8	95.4
		SM	94.9	94.8	94.5	96.0	94.5	94.1	95.1	96.6
		SH	95.1	93.9	94.9	94.7	94.0	80.2	94.9	95.1

Table 14: Average width of CI under MAR-B

Dropout rate		Variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MAR-B1	MAR-B2	MAR-B1	MAR-B2	MAR-B1	MAR-B2	MAR-B1	MAR-B2
10%	EQ	WL	11.64	11.86	11.53	11.73	11.50	11.29	11.82	12.08
		WM	11.53	11.64	11.46	11.56	11.14	10.96	11.61	11.73
		WH	11.50	11.62	11.45	11.55	11.00	10.84	11.55	11.68
		SL	11.54	11.73	11.39	11.54	11.54	11.36	11.72	11.96
		SM	11.62	11.81	11.47	11.63	11.00	10.87	11.65	11.86
		SH	11.54	11.67	11.45	11.55	10.97	10.82	11.54	11.67
	HE	WL	11.73	11.89	11.62	11.76	11.73	11.28	11.92	12.13
		WM	11.65	11.78	11.58	11.70	11.25	10.92	11.75	11.90
		WH	11.56	11.66	11.50	11.59	11.06	10.80	11.61	11.73
		SL	11.72	11.96	11.53	11.71	11.71	11.42	11.89	12.20
		SM	11.64	11.83	11.50	11.65	11.05	10.88	11.69	11.88
		SH	11.60	11.72	11.50	11.60	11.02	10.77	11.63	11.74
	HC	WL	11.73	11.89	11.62	11.76	11.50	11.30	11.97	12.09
		WM	11.65	11.79	11.58	11.71	11.13	10.97	11.77	11.88
		WH	11.55	11.65	11.49	11.59	11.00	10.85	11.60	11.71
		SL	11.73	11.97	11.53	11.73	11.49	11.35	11.93	12.22
		SM	11.64	11.84	11.51	11.66	10.98	10.88	11.68	11.85
		SH	11.60	11.72	11.50	11.59	10.95	10.81	11.60	11.74
30%	EQ	WL	13.29	14.33	12.99	13.90	11.84	11.72	13.67	14.81
		WM	13.25	14.27	13.03	13.94	11.03	10.87	13.41	14.41
		WH	13.05	13.89	12.88	13.64	10.82	10.59	13.07	13.89
		SL	12.92	13.73	12.49	13.13	11.81	11.81	13.18	14.02
		SM	12.98	13.81	12.63	13.29	10.87	10.84	12.96	13.75
		SH	13.05	13.89	12.74	13.43	10.67	10.50	12.92	13.68
	HE	WL	13.70	14.68	13.32	14.17	11.73	11.57	14.15	15.20
		WM	13.53	14.43	13.26	14.07	11.00	10.85	13.65	14.55
		WH	13.42	14.20	13.20	13.90	10.82	10.59	13.43	14.24
		SL	13.72	14.80	13.07	13.92	11.60	11.68	13.94	15.10
		SM	13.29	14.08	12.85	13.48	10.83	10.78	13.25	13.99
		SH	13.46	14.33	13.05	13.75	10.65	10.48	13.31	14.04
	HC	WL	13.71	14.71	13.32	14.21	11.55	11.56	14.08	15.21
		WM	13.56	14.40	13.29	14.04	10.93	10.85	13.72	14.57
		WH	13.41	14.20	13.19	13.90	10.79	10.62	13.43	14.20
		SL	13.69	14.80	13.05	13.91	11.51	11.64	13.94	15.06
		SM	13.35	14.06	12.89	13.46	10.75	10.74	13.29	13.96
		SH	13.49	14.32	13.07	13.74	10.61	10.49	13.35	14.07

Table 15: Coverage (%) under MAR-L

Dropout rate		variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MAR-L1	MAR-L2	MAR-L1	MAR-L2	MAR-L1	MAR-L2	MAR-L1	MAR-L2
10%	EQ	WL	95.7	94.7	95.4	95.8	97.6	94.2	96.0	96.0
		WM	94.5	93.0	94.6	95.4	96.5	92.2	95.1	96.0
		WH	95.9	92.0	95.7	95.1	96.6	78.4	95.5	95.9
		SL	95.0	92.5	94.9	95.1	97.1	96.8	95.1	95.4
		SM	95.6	92.8	95.3	94.6	96.6	95.7	95.5	94.6
		SH	94.6	91.7	94.5	94.1	95.8	90.0	94.7	94.3
	HE	WL	95.0	93.6	95.1	95.1	92.0	96.6	95.7	95.7
		WM	94.8	93.9	95.0	94.2	86.0	95.1	95.5	95.1
		WH	94.7	92.1	95.6	95.9	83.2	85.4	95.6	96.0
		SL	94.6	92.8	94.6	94.3	90.5	95.7	94.9	94.8
		SM	94.6	92.8	95.4	94.8	90.9	96.1	95.5	95.1
		SH	93.8	92.4	94.3	94.5	87.2	93.7	94.5	95.0
	HC	WL	95.2	95.4	95.2	94.9	96.2	88.4	95.3	95.7
		WM	94.3	93.8	95.4	95.4	94.5	83.5	95.3	95.5
		WH	95.4	92.2	95.7	95.8	93.4	67.2	96.2	95.9
		SL	95.8	94.9	94.9	95.2	97.7	95.0	95.6	95.7
		SM	93.8	91.8	94.1	94.9	96.0	92.9	94.2	94.9
		SH	93.1	92.5	94.9	93.6	93.9	83.6	94.9	93.6
30%	EQ	WL	96.3	93.3	95.9	94.3	96.5	93.7	96.4	95.7
		WM	95.8	89.2	95.8	96.3	93.4	83.6	96.0	97.3
		WH	96.2	81.9	95.6	93.7	92.7	30.6	95.3	94.5
		SL	95.2	91.1	94.7	94.7	95.7	98.5	95.1	95.6
		SM	94.2	86.5	94.0	93.4	93.2	97.1	94.9	94.0
		SH	95.2	80.9	94.7	94.7	92.2	81.7	94.9	94.8
	HE	WL	94.5	92.8	94.8	95.7	62.1	97.4	95.4	96.8
		WM	92.7	85.9	94.2	93.4	42.5	95.1	94.8	94.8
		WH	91.2	75.5	94.4	93.6	25.6	62.5	95.3	94.7
		SL	93.9	90.4	94.1	94.3	76.2	92.9	95.8	96.6
		SM	93.5	85.9	94.9	94.0	60.8	91.5	95.3	94.5
		SH	88.4	74.8	94.3	95.0	48.9	94.9	94.3	95.6
	HC	WL	93.9	90.9	95.3	94.7	95.9	71.9	96.1	96.2
		WM	94.0	86.1	95.1	95.0	91.5	48.1	95.9	96.3
		WH	90.6	76.2	95.3	93.9	81.6	6.9	95.2	95.0
		SL	93.4	89.9	93.7	93.6	99.1	95.1	95.2	95.7
		SM	92.4	86.5	94.5	94.3	94.7	84.8	95.2	95.3
		SH	89.4	74.3	94.6	94.3	91.2	47.8	95.2	94.9

Table 16: Average width of CI under MAR-L

Dropout rate		variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MAR-L1	MAR-L2	MAR-L1	MAR-L2	MAR-L1	MAR-L2	MAR-L1	MAR-L2
10%	EQ	WL	11.64	11.67	11.58	11.60	12.53	11.27	11.87	12.02
		WM	11.51	11.53	11.48	11.50	11.74	11.02	11.61	11.71
		WH	11.47	11.49	11.45	11.46	11.51	10.94	11.53	11.63
		SL	11.71	11.88	11.56	11.68	12.44	11.31	11.98	12.28
		SM	11.56	11.68	11.47	11.57	11.53	10.87	11.67	11.81
		SH	11.54	11.62	11.48	11.55	11.37	10.86	11.61	11.71
	HE	WL	11.65	11.73	11.59	11.66	12.67	11.06	11.92	12.04
		WM	11.67	11.70	11.63	11.66	11.91	10.83	11.83	11.91
		WH	11.50	11.54	11.48	11.51	11.54	10.81	11.59	11.66
		SL	11.64	11.77	11.51	11.63	12.50	11.22	11.85	12.08
		SM	11.58	11.69	11.50	11.60	11.59	10.80	11.68	11.84
		SH	11.55	11.61	11.50	11.54	11.40	10.71	11.61	11.70
	HC	WL	11.65	11.72	11.58	11.65	12.28	11.58	11.91	12.02
		WM	11.66	11.72	11.62	11.67	11.71	11.18	11.80	11.92
		WH	11.50	11.54	11.48	11.51	11.45	11.07	11.59	11.67
		SL	11.65	11.78	11.51	11.62	12.11	11.49	11.88	12.09
		SM	11.58	11.67	11.50	11.58	11.45	10.98	11.68	11.83
		SH	11.55	11.62	11.49	11.55	11.26	10.96	11.62	11.71
30%	EQ	WL	13.27	13.55	13.03	13.27	13.05	11.46	13.73	14.35
		WM	13.21	13.47	13.04	13.28	11.87	10.69	13.41	13.98
		WH	13.00	13.21	12.88	13.08	11.38	10.47	13.10	13.55
		SL	12.83	13.30	12.48	12.86	12.72	11.63	13.16	13.78
		SM	12.88	13.33	12.61	12.98	11.56	10.70	12.97	13.54
		SH	12.97	13.45	12.74	13.13	11.16	10.40	12.93	13.50
	HE	WL	13.48	13.73	13.21	13.43	13.13	11.10	14.08	14.67
		WM	13.41	13.64	13.23	13.44	11.96	10.47	13.71	14.27
		WH	13.24	13.42	13.10	13.27	11.47	10.32	13.43	13.87
		SL	13.52	14.18	13.02	13.57	12.82	11.30	13.98	15.09
		SM	13.07	13.51	12.76	13.12	11.63	10.48	13.21	13.80
		SH	13.28	13.72	12.99	13.36	11.23	10.22	13.29	13.92
	HC	WL	13.44	13.69	13.18	13.40	12.78	11.56	14.07	14.72
		WM	13.43	13.62	13.25	13.42	11.76	10.84	13.71	14.27
		WH	13.21	13.43	13.08	13.27	11.35	10.65	13.40	13.89
		SL	13.55	14.13	13.03	13.51	12.64	11.67	14.00	14.98
		SM	13.09	13.51	12.77	13.12	11.40	10.82	13.19	13.80
		SH	13.28	13.73	13.00	13.36	11.11	10.52	13.31	13.96

Table 17(A): Coverage (%) under MNAR in relation to dropout rate

variance-covariance	Dropout rate between groups		CCA MNAR-1	CCA MNAR-2	MMRM MNAR-1	MMRM MNAR-2	LOCF MNAR-1	LOCF MNAR-2	MI MNAR-1	MI MNAR-2
SM	EQ	10	95.0	85.3	95.3	87.9	95.4	89.9	94.7	88.3
		20	94.4	66.8	94.5	76.9	94.9	82.1	94.6	79.0
		30	94.7	61.0	94.0	73.1	92.8	75.5	94.5	75.0
		40	96.4	37.8	95.2	55.6	92.5	58.4	96.2	58.4
		50	96.2	36.7	95.8	54.4	87.8	48.8	96.8	59.3
	HE	10	93.6	83.5	93.6	87.7	94.6	84.1	94.0	87.9
		20	90.7	65.6	91.8	76.1	89.3	63.9	92.4	77.7
		30	81.3	50.7	87.0	66.9	84.5	35.8	87.8	68.6
		40	79.5	41.1	84.5	57.8	77.5	21.8	87.4	61.2
		50	80.8	30.7	85.9	51.1	76.1	15.5	87.5	56.3
	HC	10	93.9	83.6	94.5	87.6	97.0	93.6	94.8	88.1
		20	91.1	63.8	92.3	76.0	96.9	92.7	92.8	75.7
		30	82.7	52.5	87.4	69.0	97.0	93.6	88.1	71.5
		40	81.0	39.0	85.3	57.7	96.4	88.9	86.5	61.3
		50	80.7	31.0	84.7	49.2	95.0	79.4	87.6	56.0

Table 17(B): Coverage (%) under MNAR in relation to covariance pattern

Dropout rate		variance-covariance	CCA		MMRM		LOCF		MI	
	between groups		MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2
30	EQ	WL	95.9	71.7	94.9	75.8	94.9	75.6	95.5	78.9
		WM	95.7	44.1	95.6	54.1	91.8	56.4	96.0	56.6
		WH	95.1	13.6	94.6	23.0	91.2	40.6	95.3	25.3
		SL	95.0	78.0	93.9	83.7	95.2	86.9	95.4	86.1
		SM	95.2	59.6	95.0	72.7	91.6	74.4	94.5	74.3
		SH	94.5	28.8	94.2	49.4	91.0	73.1	94.5	51.4
	HE	WL	84.6	65.7	85.5	71.2	88.8	36.8	87.3	73.1
		WM	74.7	33.6	77.3	43.4	86.7	16.2	78.7	46.5
		WH	49.8	6.0	59.8	12.0	91.8	11.1	60.2	13.2
		SL	89.6	66.3	89.8	76.2	87.4	55.8	92.0	79.4
		SM	82.1	48.4	87.3	65.8	84.4	33.5	88.2	68.9
		SH	65.9	15.5	75.9	35.3	86.8	31.7	76.3	38.8
	HC	WL	85.8	61.9	87.2	68.4	98.2	94.3	89.5	73.8
		WM	71.7	33.5	76.0	41.1	95.0	86.7	78.5	44.3
		WH	48.7	5.7	59.0	11.4	91.1	74.1	60.0	12.3
		SL	90.0	68.5	91.3	77.3	98.3	95.9	92.8	81.1
		SM	83.2	50.0	87.1	67.0	97.1	93.9	88.5	70.7
		SH	66.9	16.4	76.5	38.4	95.2	91.4	77.5	40.7

Table 18(A): Average width of CI under MNAR in relation to dropout rate

variance-covariance	Dropout rate		CCA		MMRM		LOCF		MI	
	between groups	%	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2
SM	EQ	10%	11.37	11.43	11.25	11.30	11.17	10.67	11.46	11.51
		20%	11.85	12.04	11.64	11.78	11.32	10.63	11.88	12.13
		30%	12.63	12.92	12.34	12.57	11.39	10.61	12.67	13.01
		40%	13.70	14.18	13.11	13.49	11.46	10.55	13.61	14.13
		50%	15.42	16.08	14.43	14.94	11.41	10.46	15.17	16.04
	HE	10%	11.35	11.42	11.24	11.30	11.23	10.61	11.43	11.50
		20%	11.92	12.12	11.71	11.87	11.40	10.48	11.97	12.21
		30%	12.70	12.96	12.35	12.57	11.48	10.26	12.70	13.01
		40%	13.82	14.32	13.28	13.67	11.48	10.17	13.86	14.36
		50%	15.64	16.27	14.65	15.14	11.44	10.12	15.51	16.27
	HC	10%	11.35	11.44	11.24	11.32	11.08	10.76	11.44	11.54
		20%	11.94	12.13	11.72	11.87	11.20	10.73	11.99	12.20
		30%	12.70	12.96	12.36	12.57	11.21	10.74	12.75	13.08
		40%	13.87	14.33	13.31	13.69	11.26	10.72	13.92	14.41
		50%	15.67	16.27	14.67	15.12	11.32	10.66	15.56	16.29

Table 18(B): Average width of CI under MNAR in relation to covariance pattern

Dropout rate		variance-covariance	CCA		MMRM		LOCF		MI	
%	between groups		MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2
No dropout			10.80		10.80					
30%	EQ	WL	12.82	12.94	12.55	12.66	12.49	11.48	13.22	13.40
		WM	12.82	12.96	12.63	12.76	11.39	10.68	13.00	13.20
		WH	12.63	12.72	12.49	12.57	10.96	10.43	12.69	12.79
		SL	12.53	12.95	12.16	12.52	12.58	11.62	12.83	13.33
		SM	12.62	12.91	12.32	12.56	11.39	10.63	12.65	13.00
		SH	12.70	12.98	12.45	12.67	11.01	10.33	12.64	12.92
	HE	WL	12.87	12.95	12.58	12.64	12.61	10.96	13.27	13.44
		WM	12.85	12.95	12.64	12.73	11.48	10.35	13.04	13.17
		WH	12.68	12.72	12.51	12.55	11.02	10.21	12.71	12.83
		SL	13.07	13.47	12.54	12.88	12.76	11.11	13.41	14.02
		SM	12.65	12.94	12.31	12.54	11.48	10.25	12.68	13.03
		SH	12.85	13.15	12.53	12.79	11.09	10.10	12.78	13.12
	HC	WL	12.82	12.88	12.53	12.57	12.17	11.56	13.29	13.39
		WM	12.82	12.95	12.61	12.73	11.23	10.82	13.00	13.17
		WH	12.64	12.74	12.48	12.56	10.87	10.56	12.69	12.81
		SL	13.09	13.52	12.56	12.91	12.50	11.74	13.50	14.03
		SM	12.69	12.97	12.34	12.58	11.22	10.79	12.73	13.06
		SH	12.85	13.14	12.53	12.78	10.92	10.49	12.77	13.10

Table 19: Power (%) under MCAR

Dropout rate		Variance-covariance	CCA	MMRM	LOCF	MI
	between groups					
10	EQ	WL	86.4	87.1	83.8	85.0
		WM	85.4	86.5	85.5	85.6
		WH	87.3	87.5	86.1	86.0
		SL	84.6	85.3	79.3	83.9
		SM	84.8	85.6	83.6	84.4
		SH	85.5	86.2	87.5	86.5
	HE	WL	87.4	88.1	68.9	87.1
		WM	85.6	85.8	77.3	85.0
		WH	86.3	86.4	76.5	86.2
		SL	84.2	85.2	64.0	84.6
		SM	85.6	86.3	76.0	85.3
		SH	85.4	86.2	75.6	86.0
	HC	WL	87.2	87.7	93.3	87.4
		WM	84.9	85.1	88.9	84.3
		WH	87.8	87.5	92.2	86.8
		SL	83.5	84.4	90.9	83.3
		SM	85.0	85.7	90.3	85.0
		SH	85.9	86.5	92.5	85.4
30	EQ	WL	77.3	78.9	68.5	76.5
		WM	75.7	76.2	72.2	75.0
		WH	76.8	78.3	73.8	77.6
		SL	77.0	79.5	66.9	75.0
		SM	77.7	80.0	74.1	78.2
		SH	77.9	81.1	78.0	79.8
	HE	WL	73.6	76.1	27.9	72.7
		WM	75.3	76.0	37.0	75.0
		WH	74.6	75.8	39.0	74.9
		SL	73.6	75.5	29.7	71.3
		SM	78.5	79.9	40.2	77.7
		SH	77.4	79.4	44.9	78.2
	HC	WL	75.7	77.2	93.3	72.7
		WM	74.3	76.6	94.3	73.8
		WH	75.6	78.0	94.2	76.2
		SL	73.1	76.7	90.4	72.3
		SM	76.1	77.6	95.3	76.4
		SH	76.5	77.9	95.3	77.1

Table 20: Power (%) under MAR-B

Dropout rate		Variance-covariance	CCA		MMRM		LOCF		MI	
	between groups		MAR-B1	MAR-B2	MAR-B1	MAR-B2	MAR-B1	MAR-B2	MAR-B1	MAR-B2
10	EQ	WL	88.1	87.2	88.8	87.7	84.4	89.7	87.2	85.1
		WM	87.1	86.0	86.9	85.8	83.5	90.8	86.5	86.2
		WH	87.4	85.0	87.6	85.4	86.1	95.4	86.5	84.5
		SL	84.3	85.2	85.6	86.4	80.7	86.8	83.9	84.0
		SM	85.9	84.4	86.3	85.7	84.6	89.7	86.1	84.6
		SH	86.7	85.9	86.4	86.9	85.5	94.1	86.7	86.0
	HE	WL	87.0	84.9	87.6	86.0	64.0	80.2	86.1	85.4
		WM	85.2	84.1	86.0	84.6	68.6	84.1	84.7	84.1
		WH	87.2	85.9	87.1	86.6	70.6	91.7	86.9	86.4
		SL	85.1	83.0	86.0	84.1	60.4	73.3	84.4	81.3
		SM	86.1	83.4	87.1	84.3	74.6	83.0	86.6	83.2
		SH	86.5	85.4	87.1	86.0	70.1	88.3	87.1	85.6
	HC	WL	87.7	87.4	87.7	87.6	93.7	96.4	86.9	86.1
		WM	86.6	84.2	86.4	84.6	92.8	96.9	86.3	84.1
		WH	85.9	86.4	86.8	85.8	94.3	97.6	86.2	85.5
		SL	84.3	82.5	85.4	84.1	91.6	94.2	83.1	81.5
		SM	85.9	84.1	86.1	85.3	90.2	94.4	85.3	84.3
		SH	85.9	84.9	86.8	85.3	95.7	97.9	85.6	84.7
30	EQ	WL	78.4	70.5	79.0	72.0	65.9	82.3	76.1	67.9
		WM	76.9	68.2	78.1	69.9	70.7	90.4	76.0	67.1
		WH	79.8	72.9	81.1	74.8	75.0	98.2	79.5	73.3
		SL	78.9	72.3	81.2	76.2	67.1	78.0	77.4	72.1
		SM	77.8	74.0	78.9	76.3	73.4	86.0	77.4	73.6
		SH	79.3	73.2	80.0	75.8	77.0	96.7	78.4	74.6
	HE	WL	74.6	67.4	76.9	70.3	25.5	49.1	72.8	64.7
		WM	74.5	66.0	76.5	68.3	29.9	62.3	74.0	65.8
		WH	74.6	69.2	76.6	69.9	27.2	83.6	74.5	68.8
		SL	73.2	65.7	77.2	70.5	32.5	44.3	72.7	64.3
		SM	75.7	70.0	77.1	73.6	35.3	56.7	75.5	69.9
		SH	76.4	69.4	77.6	73.5	32.0	78.6	76.9	71.7
	HC	WL	75.4	68.8	76.6	70.7	94.6	97.8	72.6	66.0
		WM	74.1	69.4	75.4	71.7	96.4	98.1	73.1	68.6
		WH	74.7	69.7	75.6	70.8	96.4	99.4	75.1	70.3
		SL	71.2	64.9	76.7	70.1	92.1	96.1	70.7	64.3
		SM	76.8	72.7	79.8	76.6	95.9	97.8	77.7	73.7
		SH	75.8	70.5	77.7	73.0	96.5	99.6	75.9	71.8

Table 21: Power (%) under MAR-L

Dropout rate		variance-covariance	CCA		MMRM		LOCF		MI	
	between groups		MAR-L1	MAR-L2	MAR-L1	MAR-L2	MAR-L1	MAR-L2	MAR-L1	MAR-L2
10	EQ	WL	87.6	80.0	88.4	87.2	79.3	96.7	86.9	85.6
		WM	86.3	77.9	86.6	86.1	83.0	96.3	85.9	85.5
		WH	86.3	71.8	86.8	85.8	82.6	99.5	85.6	84.2
		SL	84.4	72.6	85.0	84.0	75.5	91.0	82.9	82.2
		SM	86.6	74.5	87.2	85.7	82.7	93.7	86.2	84.8
		SH	86.2	68.4	86.6	86.8	83.9	97.1	86.2	85.4
	HE	WL	90.9	80.8	87.6	87.3	47.6	91.1	86.0	85.3
		WM	88.0	76.1	85.0	85.2	50.6	94.8	83.1	84.3
		WH	89.9	73.4	85.2	86.5	53.7	98.8	84.7	85.6
		SL	87.5	76.3	85.8	85.0	52.2	81.0	84.5	83.2
		SM	88.6	72.9	85.2	85.4	66.1	89.9	84.5	84.9
		SH	90.6	72.5	86.8	87.6	60.8	94.2	86.3	86.5
	HC	WL	84.7	80.7	88.5	87.5	95.4	98.5	86.7	86.2
		WM	80.4	75.3	85.8	85.7	96.7	99.4	85.3	84.4
		WH	79.4	72.1	85.4	86.1	97.0	99.9	84.9	85.4
		SL	82.3	75.6	86.2	85.2	91.6	96.1	84.3	82.7
		SM	82.8	72.2	86.9	84.2	91.3	96.1	85.8	82.8
		SH	81.1	73.2	86.4	86.4	96.0	99.2	85.9	85.9
30	EQ	WL	78.4	58.6	80.0	78.7	59.8	97.2	75.9	72.2
		WM	77.2	48.2	78.3	75.7	66.1	99.6	77.5	70.8
		WH	77.2	34.4	77.7	76.9	69.4	100.0	77.0	73.9
		SL	77.6	57.5	79.8	77.5	61.7	87.8	76.7	72.9
		SM	77.9	45.7	79.8	75.5	69.3	94.9	77.8	73.5
		SH	78.6	34.7	80.7	76.9	74.6	99.6	79.7	75.5
	HE	WL	86.1	52.8	78.5	75.3	9.2	85.0	73.9	68.1
		WM	85.6	41.3	74.6	74.2	9.5	95.0	71.9	69.9
		WH	90.6	26.4	77.0	74.6	4.5	99.9	74.5	71.1
		SL	82.4	44.2	77.3	73.5	23.8	72.8	71.7	67.5
		SM	87.3	43.7	76.1	74.4	21.3	78.7	73.9	71.1
		SH	91.2	25.6	77.4	76.9	18.5	97.3	75.2	74.1
	HC	WL	64.0	54.8	77.1	76.6	96.3	99.7	71.7	70.2
		WM	58.7	41.9	76.5	74.2	98.8	99.9	73.8	70.0
		WH	50.0	25.7	75.5	75.8	99.6	100.0	73.7	71.9
		SL	62.4	43.3	75.1	71.9	85.8	97.2	70.2	64.7
		SM	60.4	42.8	79.5	77.7	95.9	99.6	76.6	73.9
		SH	51.0	25.4	76.8	75.9	98.2	100.0	74.3	72.4

Table 22(A): Power (%) under MNAR in relation to dropout rate

variance-covariance	Dropout rate		CCA		MMRM		LOCF		MI	
	between groups	%	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2
SM	EQ	10	86.7	56.8	87.1	66.9	82.9	76.5	86.2	66.3
		20	84.3	29.2	84.8	44.3	79.2	63.0	83.8	42.2
		30	80.4	17.9	81.4	32.1	70.4	51.6	80.2	31.0
		40	74.9	4.7	77.9	12.6	66.7	33.2	75.3	11.0
		50	64.8	4.7	69.2	8.8	58.5	26.4	65.0	7.1
	HE	10	92.4	56.4	91.7	67.3	79.4	66.9	91.2	66.1
		20	93.2	27.5	91.6	41.2	66.0	41.0	90.8	40.1
		30	96.6	11.7	94.6	27.3	52.7	18.7	94.2	25.6
		40	95.7	4.3	93.3	12.5	41.3	10.7	92.7	11.0
		50	92.2	5.5	90.4	6.7	35.3	6.6	87.9	5.0
	HC	10	79.8	57.1	82.0	66.1	87.1	83.2	81.0	64.0
		20	67.2	26.8	73.9	42.1	86.7	77.9	72.6	39.4
		30	42.4	13.5	56.4	27.4	86.3	81.1	53.3	24.8
		40	32.2	5.9	44.5	12.8	82.5	71.9	41.6	12.0
		50	17.8	5.6	31.4	7.0	74.9	55.0	26.6	5.0

Table 22(B): Power (%) under MNAR in relation to covariance pattern

Dropout rate		variance-covariance	CCA		MMRM		LOCF		MI	
	between groups		MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2	MNAR-1	MNAR-2
30	EQ	WL	79.8	25.5	81.1	32.1	61.9	38.9	77.8	28.4
		WM	78.9	10.1	79.8	14.6	68.2	28.6	77.5	13.9
		WH	80.6	4.9	80.4	5.6	73.8	22.5	79.3	5.2
		SL	76.4	35.1	78.9	49.9	61.2	53.7	75.9	45.1
		SM	78.9	17.0	80.8	31.4	67.9	51.1	79.3	29.4
		SH	79.6	5.3	81.3	13.4	75.3	54.7	79.8	12.7
	HE	WL	96.2	20.6	95.6	27.9	41.1	12.5	94.5	25.0
		WM	98.1	6.4	97.6	10.0	58.2	6.0	97.0	8.9
		WH	99.7	12.1	99.5	6.7	74.5	6.0	99.5	5.6
		SL	91.3	18.8	89.1	34.0	40.9	21.7	87.1	28.6
		SM	96.7	12.2	94.8	24.9	51.9	19.4	94.0	22.8
		SH	98.3	5.6	98.0	7.4	60.3	18.7	97.7	7.7
	HC	WL	45.8	19.1	52.0	27.2	82.4	74.9	47.3	23.1
		WM	29.6	5.2	34.7	8.7	80.3	63.8	33.2	8.1
		WH	12.6	10.2	18.6	4.6	74.7	50.8	17.3	4.4
		SL	51.9	19.8	61.6	34.2	76.7	73.7	56.4	28.5
		SM	41.4	12.3	55.5	25.5	87.8	79.7	52.2	24.3
		SH	24.3	6.3	40.7	8.0	84.2	79.1	38.5	7.5

Appendix 6: MMRM versus cLDA (tables 23 and 24)

Table 23: Effects of baseline handling strategies in a repeated measurement analysis model on overall accuracy

Patterns under 30% dropout rate		Bias				RMSE			
		With Kenward-Roger correction		Without Kenward-Roger correction		With Kenward-Roger correction		Without Kenward-Roger correction	
		MMRM	cLDA	MMRM	cLDA	MMRM	cLDA	MMRM	cLDA
No missing		-0.044	-0.044	-0.044	-0.044	2.338	2.338	2.338	2.338
Equal dropout rate between groups	MCAR	0.005	0.005	0.005	0.005	2.549	2.549	2.549	2.549
	MAR-B1	0.015	0.015	0.015	0.015	2.520	2.520	2.520	2.520
	MAR-B2	0.113	0.113	0.113	0.113	2.601	2.601	2.601	2.601
	MAR-L1	-0.029	-0.029	-0.029	-0.029	2.514	2.514	2.514	2.514
	MAR-L2	-0.038	-0.038	-0.038	-0.038	2.611	2.611	2.611	2.611
	MNAR-1	-0.137	-0.137	-0.137	-0.137	2.496	2.496	2.496	2.496
	MNAR-2	4.240	4.240	4.240	4.240	4.936	4.936	4.936	4.936
High dropout rate in the experimental group	MCAR	-0.047	-0.047	-0.047	-0.047	2.546	2.546	2.546	2.546
	MAR-B1	-0.073	-0.073	-0.073	-0.073	2.564	2.564	2.564	2.564
	MAR-B2	-0.028	-0.028	-0.028	-0.028	2.638	2.638	2.638	2.638
	MAR-L1	-0.017	-0.017	-0.017	-0.017	2.579	2.579	2.579	2.579
	MAR-L2	-0.113	-0.113	-0.113	-0.113	2.621	2.621	2.621	2.621
	MNAR-1	-2.522	-2.522	-2.522	-2.522	3.552	3.552	3.552	3.552
	MNAR-2	4.800	4.800	4.800	4.800	5.429	5.429	5.429	5.429
High dropout rate in the control group	MCAR	-0.006	-0.006	-0.006	-0.006	2.550	2.550	2.550	2.550
	MAR-B1	-0.028	-0.028	-0.028	-0.028	2.559	2.559	2.559	2.559
	MAR-B2	0.083	0.083	0.083	0.083	2.619	2.619	2.619	2.619
	MAR-L1	0.126	0.126	0.126	0.126	2.559	2.559	2.559	2.559
	MAR-L2	-0.014	-0.014	-0.014	-0.014	2.599	2.599	2.599	2.599
	MNAR-1	2.539	2.539	2.539	2.539	3.574	3.574	3.574	3.574
	MNAR-2	4.871	4.871	4.871	4.871	5.499	5.499	5.499	5.499

Table 24: Effects of baseline handling strategies in a repeated measurement analysis model on coverage and power

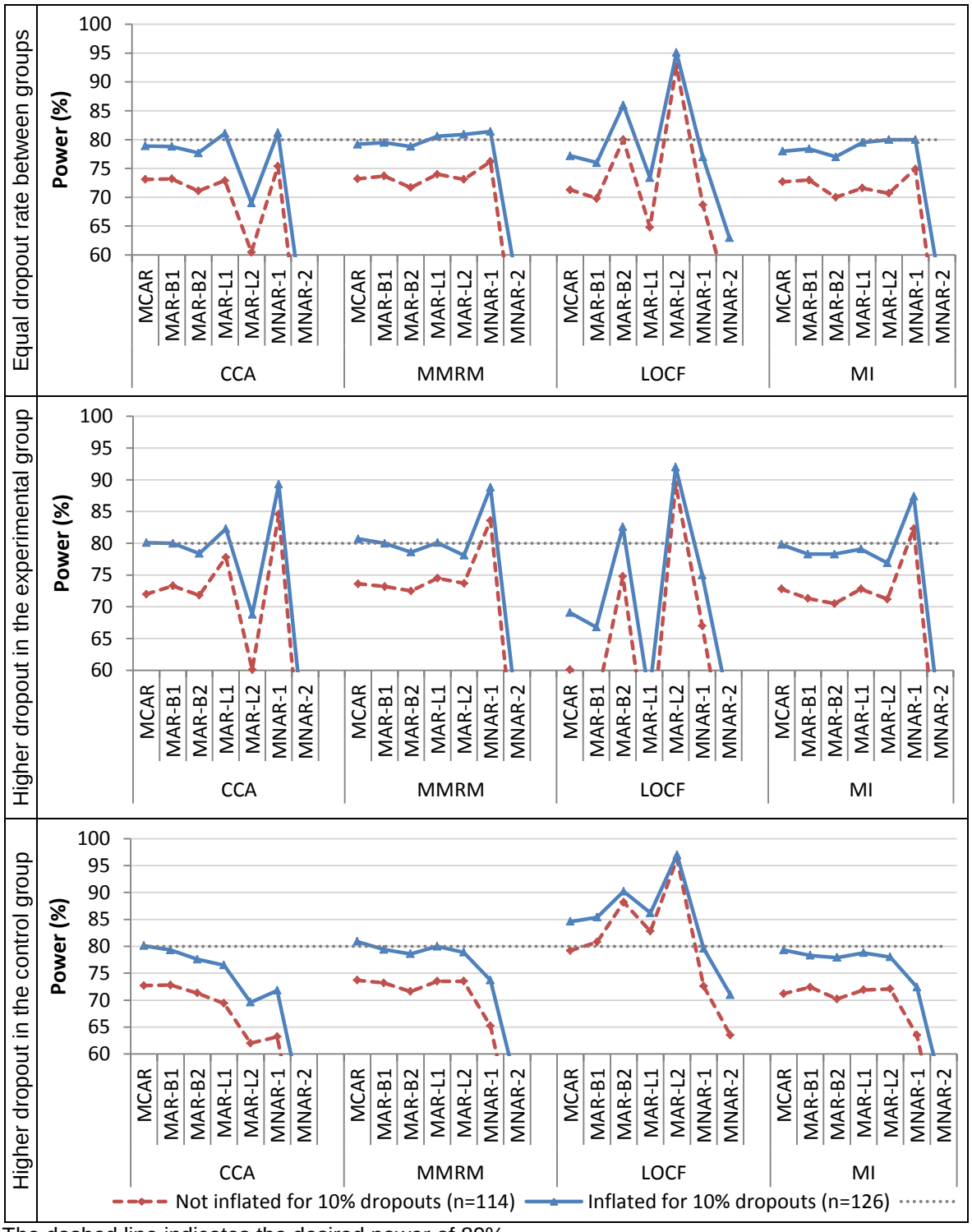
Patterns under 30% dropout rate		Coverage				Observed power			
		With Kenward-Roger correction		Without Kenward-Roger correction		With Kenward-Roger correction		Without Kenward-Roger correction	
		MMRM	cLDA	MMRM	cLDA	MMRM	cLDA	MMRM	cLDA
No missing		95.0%	95.0%	95.0%	94.8%	90.5%	90.7%	90.6%	90.8%
Equal dropout rate between groups	MCAR	94.3%	94.3%	94.1%	93.7%	78.9%	78.8%	79.1%	80.0%
	MAR-B1	95.6%	95.6%	95.6%	95.4%	78.4%	78.1%	78.7%	79.5%
	MAR-B2	94.6%	93.2%	94.3%	92.9%	75.3%	78.0%	75.5%	79.1%
	MAR-L1	95.0%	95.0%	94.9%	94.6%	80.3%	79.8%	80.5%	81.4%
	MAR-L2	94.3%	93.3%	94.2%	93.0%	76.6%	78.4%	77.0%	79.5%
	MNAR-1	94.8%	94.8%	94.6%	94.2%	82.3%	82.4%	82.6%	83.5%
	MNAR-2	73.1%	71.8%	72.8%	70.9%	31.2%	32.3%	32.0%	33.6%
High dropout rate in the experimental group	MCAR	95.1%	95.0%	94.9%	94.4%	79.2%	78.9%	79.5%	80.2%
	MAR-B1	95.5%	95.2%	95.4%	94.7%	77.3%	77.5%	77.7%	79.1%
	MAR-B2	94.7%	93.2%	94.6%	92.4%	74.8%	79.0%	75.6%	80.3%
	MAR-L1	94.3%	94.2%	94.2%	93.9%	77.5%	77.4%	78.0%	78.7%
	MAR-L2	94.7%	93.9%	94.6%	93.2%	76.2%	78.2%	76.8%	79.4%
	MNAR-1	87.6%	87.4%	87.2%	86.7%	95.2%	95.3%	95.6%	95.8%
	MNAR-2	67.6%	66.1%	66.7%	64.9%	25.7%	27.1%	26.2%	28.0%
High dropout rate in the control group	MCAR	95.0%	94.8%	94.8%	94.4%	79.3%	79.2%	79.5%	80.1%
	MAR-B1	94.3%	94.3%	94.3%	93.8%	79.3%	79.4%	79.6%	80.7%
	MAR-B2	94.9%	93.8%	94.8%	93.2%	73.6%	77.5%	74.2%	78.2%
	MAR-L1	94.1%	93.9%	93.9%	93.4%	77.0%	77.4%	77.9%	78.9%
	MAR-L2	94.9%	94.1%	94.5%	93.1%	76.7%	78.1%	77.0%	78.7%
	MNAR-1	86.9%	86.7%	86.9%	86.4%	53.8%	53.7%	54.2%	54.9%
	MNAR-2	67.4%	66.7%	67.1%	65.0%	25.5%	26.7%	26.0%	27.7%

Appendix 7: Effect of sample size on statistical power (tables 25 and 26; figures 1 and 2)

Table 25: The observed power (%) with inflated sample size – (the desired power was 90% under a strong correlation matrix)

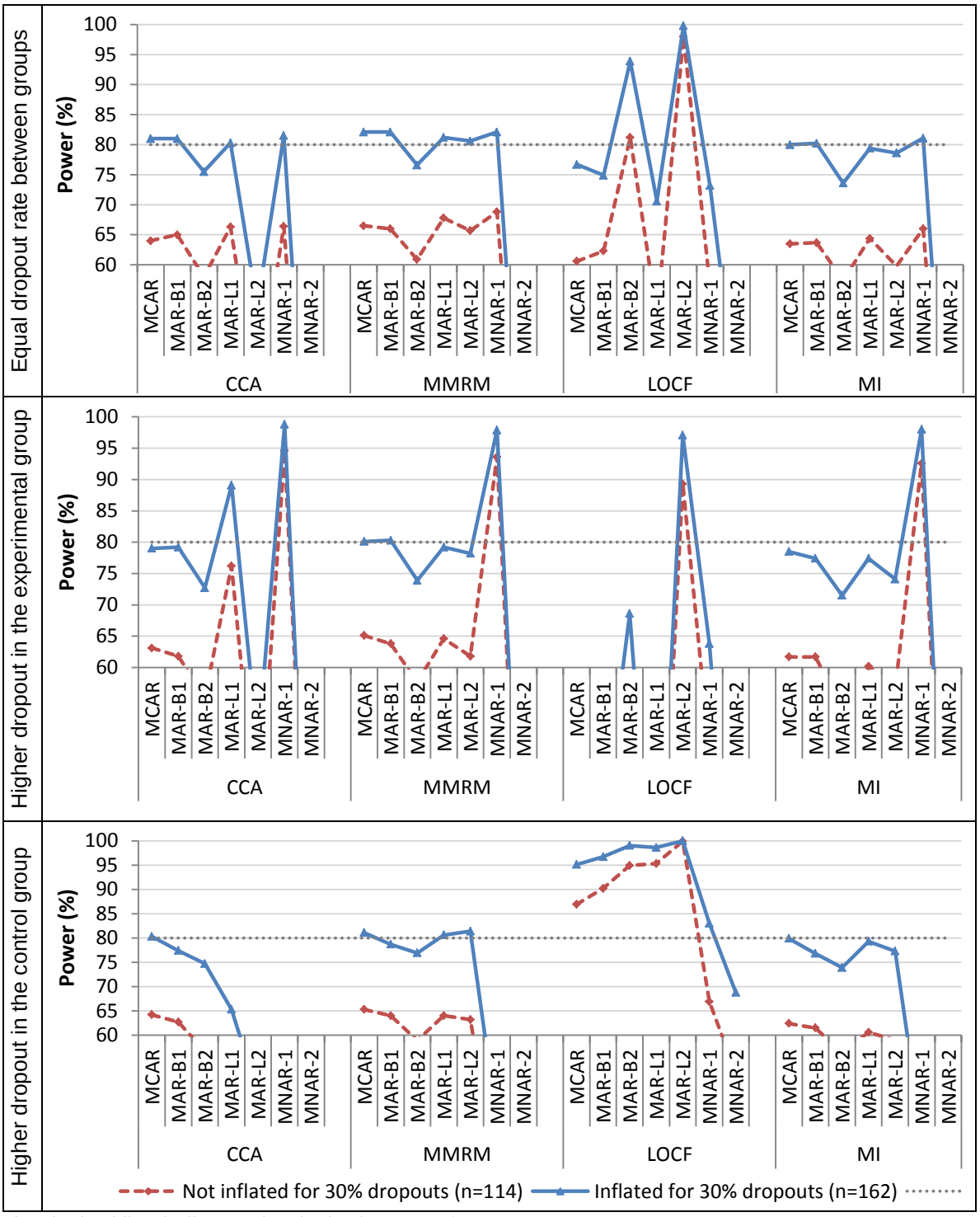
Mechanism	10% dropouts (n = 132)			30% dropouts (n = 168)		
	CCA	MMRM	MI	CCA	MMRM	MI
Equal dropout between groups						
MCAR	88.8	89.3	88.0	88.6	90.3	89.5
MAR-B1	89.4	89.8	89.2	89.9	91.7	90.4
MAR-B2	89.0	89.8	88.4	84.1	89.0	86.4
MAR-L1	89.3	89.2	88.9	90.4	92.1	90.8
MAR-L2	78.0	89.0	88.4	55.9	88.3	86.4
MNAR-1	90.4	90.7	89.1	92.0	93.1	91.9
MNAR-2	62.3	70.9	69.8	17.7	34.8	32.7
Higher dropout in the experimental group						
MCAR	89.2	89.8	89.3	88.1	90.2	89.0
MAR-B1	89.3	90.0	89.3	88.9	90.7	90.1
MAR-B2	87.7	88.6	88.2	84.2	88.1	86.4
MAR-L1	92.8	90.0	89.3	96.2	89.3	87.8
MAR-L2	76.7	89.8	88.7	49.9	87.7	85.6
MNAR-1	95.6	95.5	95.5	99.7	98.4	98.2
MNAR-2	59.4	67.6	67.5	13.1	28.0	26.5
Higher dropout in the control group						
MCAR	89.7	90.6	89.0	87.4	89.4	88.7
MAR-B1	88.6	89.4	88.0	88.9	90.8	90.0
MAR-B2	89.1	89.6	88.3	84.1	85.6	84.5
MAR-L1	85.2	89.2	88.4	74.6	91.1	89.6
MAR-L2	77.5	89.9	89.5	51.5	88.7	86.3
MNAR-1	79.3	83.0	82.4	54.2	68.6	65.9
MNAR-2	57.7	66.8	64.4	11.1	29.4	25.1

Figure 1: Statistical power under different sample sizes (10% dropouts and 80% desired power under a weak correlation matrix)



The dashed line indicates the desired power of 80%

Figure 2: Statistical power under different sample sizes (30% dropouts and 80% desired power under a weak correlation matrix)



The dashed line indicates the desired power of 80%

Table 26: The observed power (%) with inflated sample size – (the desired power was 80% under a strong correlation matrix)

Mechanism	10% dropouts (n = 98)			30% dropouts (n = 126)		
	CCA	MMRM	MI	CCA	MMRM	MI
Equal dropout between groups						
MCAR	81.8	81.8	81.2	76.0	79.3	77.2
MAR-B1	81.0	81.9	81.5	79.0	80.9	78.9
MAR-B2	79.6	80.0	79.6	72.7	76.2	72.9
MAR-L1	81.1	81.2	80.8	78.0	79.5	78.4
MAR-L2	71.6	80.5	79.7	41.4	75.5	71.5
MNAR-1	82.4	82.9	82.1	81.1	83.1	80.9
MNAR-2	58.5	65.2	63.7	11.6	26.0	23.7
Higher dropout in the experimental group						
MCAR	79.6	81.6	80.2	77.6	79.9	77.4
MAR-B1	81.0	80.8	80.4	76.9	78.9	77.2
MAR-B2	81.2	81.5	81.2	70.7	73.5	71.7
MAR-L1	83.7	80.7	79.8	88.7	79.3	75.5
MAR-L2	72.6	81.7	80.7	39.0	75.3	71.4
MNAR-1	89.1	87.9	86.9	97.3	96.2	95.5
MNAR-2	58.4	65.9	64.3	6.8	18.9	17.0
Higher dropout in the control group						
MCAR	81.8	81.3	80.4	76.2	80.3	78.1
MAR-B1	80.2	81.5	80.4	75.9	78.9	76.5
MAR-B2	79.5	81.3	80.2	72.9	75.4	71.5
MAR-L1	76.9	82.2	81.3	61.1	80.2	76.8
MAR-L2	71.1	82.1	81.4	38.1	76.7	72.7
MNAR-1	71.8	74.9	74.7	40.8	52.5	49.8
MNAR-2	57.0	64.7	63.4	7.9	20.6	17.8

Appendix 8: TATE trial: MI-inclusive imputation models (tables 27–30)

Table 27: Inclusive model for pain intensity at month 12

model	b3	se3	p3	Effect size
1	-0.458	0.305	0.136	-0.223
2	-0.459	0.302	0.130	-0.223
3	-0.449	0.306	0.143	-0.219
4	-0.456	0.306	0.138	-0.222
5	-0.489	0.308	0.114	-0.238
6	-0.489	0.309	0.115	-0.238
7	-0.462	0.302	0.128	-0.225
8	-0.497	0.306	0.106	-0.242

model

1 mi impute chained (regress) w6_pain m6_pain m12_pain = b_pain group1 Age Sex, add(500) burnin(50)

2 mi impute chained (regress) w6_pain m6_pain m12_pain w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot = b_PRTEE_tot b_pain group1 Age Sex, add(500) burnin(50)

3 mi impute chained (regress) w6_pain m6_pain m12_pain w6_SF_PCS m6_SF_PCS m12_SF_PCS = b_SF_PCS b_pain group1 Age Sex, add(500) burnin(50)

4 mi impute chained (regress) w6_pain m6_pain m12_pain w6_SF_MCS m6_SF_MCS m12_SF_MCS = b_SF_MCS b_pain group1 Age Sex, add(500) burnin(50)

5 mi impute chained (regress) w6_pain m6_pain m12_pain w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_SF_PCS m6_SF_PCS m12_SF_PCS = b_SF_PCS b_PRTEE_tot b_pain group1 Age Sex, add(500) burnin(50)

6 mi impute chained (regress) w6_pain m6_pain m12_pain w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_SF_MCS m6_SF_MCS m12_SF_MCS= b_SF_MCS b_PRTEE_tot b_pain group1 Age Sex, add(500) burnin(50)

7 mi impute chained (regress) w6_pain m6_pain m12_pain w6_SF_MCS m6_SF_MCS m12_SF_MCS w6_SF_PCS m6_SF_PCS m12_SF_PCS = b_SF_PCS b_SF_MCS b_pain group1 Age Sex, add(500) burnin(50)

8 mi impute chained (regress) w6_pain m6_pain m12_pain w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_SF_PCS m6_SF_PCS m12_SF_PCS w6_SF_MCS m6_SF_MCS m12_SF_MCS = b_SF_MCS b_SF_PCS b_PRTEE_tot b_pain group1 Age Sex, add(500) burnin(50)

Table 28: Inclusive model for PRTEE at month 12

model	b3	se3	p3	Effect size
1	-3.817	3.038	0.211	-0.215
2	-3.427	2.742	0.213	-0.193
3	-4.620	2.991	0.125	-0.261
4	-3.891	2.961	0.191	-0.219
5	-3.760	2.822	0.185	-0.212
6	-3.573	2.836	0.210	-0.202
7	-4.707	2.979	0.116	-0.265
8	-3.980	2.769	0.153	-0.224

model

1 mi impute chained (regress) w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot = b_PRTEE_tot b_pain group1 Age Sex, add(500) burnin(50)

2 mi impute chained (regress) w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_pain m6_pain m12_pain = b_pain b_PRTEE_tot group1 Age Sex, add(500) burnin(50)

3 mi impute chained (regress) w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_SF_PCS m6_SF_PCS m12_SF_PCS = b_pain b_SF_PCS b_PRTEE_tot group1 Age Sex, add(500) burnin(50)

4 mi impute chained (regress) w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_SF_MCS m6_SF_MCS m12_SF_MCS = b_pain b_SF_MCS b_PRTEE_tot group1 Age Sex, add(500) burnin(50)

5 mi impute chained (regress) w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_pain m6_pain m12_pain w6_SF_PCS m6_SF_PCS m12_SF_PCS = b_SF_PCS b_pain b_PRTEE_tot group1 Age Sex, add(500) burnin(50)

6 mi impute chained (regress) w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_pain m6_pain m12_pain w6_SF_MCS m6_SF_MCS m12_SF_MCS= b_SF_MCS b_pain b_PRTEE_tot group1 Age Sex, add(500) burnin(50)

7 mi impute chained (regress) w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_SF_MCS m6_SF_MCS m12_SF_MCS w6_SF_PCS m6_SF_PCS m12_SF_PCS = b_pain b_SF_PCS b_SF_MCS b_PRTEE_tot group1 Age Sex, add(500) burnin(50)

8 mi impute chained (regress) w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_pain m6_pain m12_pain w6_SF_PCS m6_SF_PCS m12_SF_PCS w6_SF_MCS m6_SF_MCS m12_SF_MCS = b_SF_MCS b_SF_PCS b_pain b_PRTEE_tot group1 Age Sex, add(500) burnin(50)

Table 29: Inclusive model for SF12-MCS at month 12

model	b3	se3	p3	Effect size
1	2.411	1.541	0.120	0.250
2	2.396	1.552	0.125	0.249
3	2.602	1.509	0.087	0.270
4	2.575	1.603	0.111	0.267
5	2.567	1.585	0.108	0.266
6	2.486	1.577	0.117	0.258
7	2.882	1.550	0.065	0.299
8	2.812	1.634	0.088	0.292

model

1 mi impute chained (regress) w6_SF_MCS m6_SF_MCS m12_SF_MCS = b_SF_MCS b_pain group1 Age Sex, add(500) burnin(50)

2 mi impute chained (regress) w6_SF_MCS m6_SF_MCS m12_SF_MCS w6_pain m6_pain m12_pain = b_pain b_SF_MCS group1 Age Sex, add(500) burnin(50)

3 mi impute chained (regress) w6_SF_MCS m6_SF_MCS m12_SF_MCS w6_SF_PCS m6_SF_PCS m12_SF_PCS = b_pain b_SF_PCS b_SF_MCS group1 Age Sex, add(500) burnin(50)

4 mi impute chained (regress) w6_SF_MCS m6_SF_MCS m12_SF_MCS w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot = b_pain b_PRTEE_tot b_SF_MCS group1 Age Sex, add(500) burnin(50)

5 mi impute chained (regress) w6_SF_MCS m6_SF_MCS m12_SF_MCS w6_pain m6_pain m12_pain w6_SF_PCS m6_SF_PCS m12_SF_PCS = b_SF_PCS b_pain b_SF_MCS group1 Age Sex, add(500) burnin(50)

6 mi impute chained (regress) w6_SF_MCS m6_SF_MCS m12_SF_MCS w6_pain m6_pain m12_pain w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot= b_PRTEE_tot b_pain b_SF_MCS group1 Age Sex, add(500) burnin(50)

7 mi impute chained (regress) w6_SF_MCS m6_SF_MCS m12_SF_MCS w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_SF_PCS m6_SF_PCS m12_SF_PCS = b_pain b_SF_PCS b_PRTEE_tot b_SF_MCS group1 Age Sex, add(500) burnin(50)

8 mi impute chained (regress) w6_SF_MCS m6_SF_MCS m12_SF_MCS w6_pain m6_pain m12_pain w6_SF_PCS m6_SF_PCS m12_SF_PCS w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot = b_PRTEE_tot b_SF_PCS b_pain b_SF_MCS group1 Age Sex, add(500) burnin(50)

Table 30: Inclusive model for SF12-PCS at month 12

model	b3	se3	p3	Effect size
1	-0.276	1.448	0.849	-0.025
2	0.025	1.451	0.987	0.002
3	-0.380	1.421	0.790	-0.035
4	-0.373	1.509	0.805	-0.034
5	-0.138	1.484	0.926	-0.013
6	-0.333	1.472	0.821	-0.030
7	-0.491	1.456	0.737	-0.045
8	-0.421	1.501	0.779	-0.038

model

1 mi impute chained (regress) w6_SF_PCS m6_SF_PCS m12_SF_PCS = b_SF_PCS b_pain group1 Age Sex, add(500) burnin(50)

2 mi impute chained (regress) w6_SF_PCS m6_SF_PCS m12_SF_PCS w6_pain m6_pain m12_pain = b_pain b_SF_PCS group1 Age Sex, add(500) burnin(50)

3 mi impute chained (regress) w6_SF_PCS m6_SF_PCS m12_SF_PCS w6_SF_MCS m6_SF_MCS m12_SF_MCS = b_pain b_SF_MCS b_SF_PCS group1 Age Sex, add(500) burnin(50)

4 mi impute chained (regress) w6_SF_PCS m6_SF_PCS m12_SF_PCS w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot = b_pain b_PRTEE_tot b_SF_PCS group1 Age Sex, add(500) burnin(50)

5 mi impute chained (regress) w6_SF_PCS m6_SF_PCS m12_SF_PCS w6_pain m6_pain m12_pain w6_SF_MCS m6_SF_MCS m12_SF_MCS = b_SF_MCS b_pain b_SF_PCS group1 Age Sex, add(500) burnin(50)

6 mi impute chained (regress) w6_SF_PCS m6_SF_PCS m12_SF_PCS w6_pain m6_pain m12_pain w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot= b_PRTEE_tot b_pain b_SF_PCS group1 Age Sex, add(500) burnin(50)

7 mi impute chained (regress) w6_SF_PCS m6_SF_PCS m12_SF_PCS w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot w6_SF_MCS m6_SF_MCS m12_SF_MCS = b_pain b_SF_MCS b_PRTEE_tot b_SF_PCS group1 Age Sex, add(500) burnin(50)

8 mi impute chained (regress) w6_SF_PCS m6_SF_PCS m12_SF_PCS w6_pain m6_pain m12_pain w6_SF_MCS m6_SF_MCS m12_SF_MCS w6_PRTEE_tot m6_PRTEE_tot m12_PRTEE_tot = b_PRTEE_tot b_SF_MCS b_pain b_SF_PCS group1 Age Sex, add(500) burnin(50)

Appendix 9: TATE results (MMRM analysis of pain intensity score)

Table 31: MMRM estimate of treatment effect from different datasets that were created by different cut-offs for reminder responses

Dataset	Description of 'missing' data	Response rate	estimate	SE	Standardized effect size	Deviation in effect size from actual data
Actual	Actual missing responses	74%	−0.456	0.299	−0.222	-
Modified 1	Actual missing and MDC responses	62%	−0.524	0.327	−0.255	0.033
Modified 2	Actual missing and 2nd & 3rd reminder responses	67%	-0.393	0.298	-0.189	-0.033
Modified 3	Actual missing, MDC and reminder (2nd & 3rd) responses	54%	−0.445	0.329	−0.217	−0.005